Atlas-Based Methods in Radiotherapy Treatment of Head and Neck Cancer

Albert K. Hoang Duc

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Centre for Medical Image Computing

University College London

August 21, 2015

I, Albert K. Hoang Duc, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Kính tặng bố tôi, người dạy tôi tất cả mọi thứ

Abstract

Radiotherapy is one of the principal methods for treating head and neck cancer (HNC). It plays an important role in the curative and palliative treatment of HNC. It uses high-energy radiation beams to kill cancer cells by damaging their DNA. Radiotherapy planning depends upon complex algorithms to determine the best trajectories and intensities of those beams by simulating their effects passing through designated areas. This requires accurate segmentation of anatomical structures and knowledge of the relative electron density within a patient body.

Computed tomography (CT) has been the modality of choice in radiotherapy planning. It offers a wealth of anatomical information and is critical in providing information about the relative electron density of tissues required to calculate radiation deposited at any one site. Manual segmentation is time-consuming and is becoming impractical with the increasing demand in image acquisition for planning. Recently, planning solely based on magnetic resonance (MR) imaging has gained popularity as it provides superior soft tissue contrast compared to CT imaging and can better facilitate the process of segmentation. However, MR imaging does not provide electron density information for dose calculation.

With the growing volumes of data and data repositories, algorithms based on atlases have gained popularity as they provide prior information for structure segmentation and tissue classification. In this PhD thesis, I demonstrate that atlas-based methods can be used for segmenting head and neck structures giving results as comparable as manual segmentation. In addition, I demonstrate that those methods can be used to support radiotherapy treatment solely based on MR imaging by generating synthetic CT images. The radiation doses calculated from a synthetic and real CT image agreed well, showing the clinical feasibility of methods based on atlases. In conclusion, I show that atlas-based methods are clinically relevant in radiotherapy treatment.

Acknowledgements

I would like to thank Prof. Ourselin from UCL and Dr. Kadir from Mirada Medical UK for taking me under their supervision.

In addition, my gratitude goes to all the people at the Center for Medical Image Computing, in particular Ron and Dominique, and to my research collaborators especially Catarina, Gemma, Jamie, Kelvin K. and Marc who provided me with valuable support.

Also, I would like to thank Abigail, Andrew, Carole, Eliza, Jonas, Marco, Maria A., Nicolas, Ninon and Sjoerd who contributed to the great work environment within the Translational Imaging Group.

Finally, I am utmost thankful to Alexander F., Gergely and Miklos for their help, kindness and friendship during the course of my PhD and beyond.

A very special thanks to Hang M.

Publication List

- **Hoang Duc A.K.**, Eminowicz G., Mendes R., Wong S.L., McClelland J., Modat M., Cardoso M.J., Mendelson A.F., Veiga C., Kadir T. and Ourselin S.: Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. Medical Physics. *In press.*

- **Hoang Duc A.K.**, Modat M., Leung K.K., Cardoso M.J., Barnes J., Kadir T., and Ourselin S. for The Alzheimers Disease Neuroimaging Initiative: Using Manifold Learning for Atlas Selection in Multi-Atlas Segmentation. (2013). PLoS ONE 8(8).

- **Hoang Duc A.K.**, Modat M., Leung K.K., Kadir T., and Ourselin S.: Manifold Learning for Atlas Selection in Multi-Atlas Based Segmentation of Hippocampus. (2012). SPIE.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

According to figures from Cancer Research UK [1], more than 331,000 people in the UK were diagnosed with cancer in 2011, and around 159,000 people died from cancer that same year. It is predicted that 30 to 40% of the population will develop some form of cancer during their lifetime. Head and neck cancer (HNC) is one type of cancer that has increased by 7% in the last 15 years making it an increased social and economic burden. The incidence among the population of HNC is 11.2 per 100,000 people. The main aetological factors of HNC are alcohol and tobacco consumption and infection with human papilloma virus. With the improvement in medical care and advances in medical technologies, half of people diagnosed with HNC now survive their disease for at least ten years and death rate has fallen by 10% over the last decade.

Radiotherapy (RT), also called radiation therapy, is one of the 3 principal methods for treating HNC alongside surgery and chemotherapy. Approximately 40% of patients will undergo radiotherapy at some time during the course of their illness. Among those patients, around 60% will be treated curatively often in combination with surgery and chemotherapy. In addition, radiotherapy plays an important role in the palliation of symptoms from HNC. It is the most suitable method for the radical treatment of localized HNC in their early stages with high-success rates where there had been no metastatic spread. Radiotherapy treatment involves the use of radiation beams of high-energy X-rays to destroy cancer cells. The effectiveness of radiotherapy ultimately depends upon the ability of complex computer algorithms to simulate the effect of those beams passing through a designated area and the amount of radiation deposited at any one site (Fraass et al., 1998). The best trajectories and intensities of the radiation beams are determined by numerous factors among which accurate segmentation of anatomical structures (Stapleford et al., 2010; Teguh et al., 2011) and the knowledge of the relative electron density within a patient body are crucial (Skrzyński et al., 2010). Indeed, volumes to be treated need to be accurately segmented in order to deliver maximum radiation dose to tumour cells while minimizing dose to critical structures. In turn, the relative electron density determines the amount of radiation absorbed by tissues and subsequently the optimal beam set up.

In radiotherapy planning, computed tomography (CT) is currently the modality of choice for clinical assessment, treatment and follow up. CT images offer a wealth of information about normal and diseased anatomy, and is critical at several stages of the radiotherapy treatment process. It also provide

[1] *www.cancerresearchuk.org*

information to determine the relative electron density required to calculate dose distribution. The acquisition of ever-increasing quantities of data has rendered manual segmentation of anatomical structures by a trained human operator impractical in a clinical routine. Manual segmentation is widely considered the gold standard but is time consuming and is subject to inter- and intra-observer variability. Recently, radiotherapy planning solely based on magnetic resonance (MR) imaging has gained popularity. MR imaging provides superior soft tissue contrast in the head and neck region (Evans, 2008) compared to CT imaging and can facilitate the process of segmentation. However, MR imaging does not provide electron-density information for dose calculation. Simple strategies consist for instance in assigning a single density value for a whole anatomical region segmented on the MR image (Karotki et al., 2011). Such method enables one to obtain a synthetic CT that can be used for dose calculation.

The result of growing volumes of data and data repositories have lead to the development of various automatic algorithms. Algorithm based on atlases have gained popularity in medical imaging as they provide prior information for structure segmentation and tissue classification. They have been mainly developed for segmenting brain structures on MR images (Artaechevarria et al., 2009; Sabuncu et al., 2010; Warfield et al., 2004). They have shown promising results, however, little work has been done in translating those algorithms to be applied in head and neck radiotherapy. As a result, the aim of this thesis is to tackle the problem of segmenting head and neck structures and estimating the relative electron density of tissues from MR images for radiotherapy planning using atlas-based algorithms.

1.1 Radiotherapy: an overview

Radiotherapy is an efficient method for treating HNC where ionizing radiation is used to eradicate malignant tumour cells or to slow down their growth. There are two ways of delivering ionizing radiation to tumour cells. First, brachytherapy or internal radiotherapy makes use of medication containing radioactive materials which are injected, either temporarily or permanently, into a vein or a body cavity near the treatment area. Second, teletherapy or external beam radiotherapy uses radiation beams of high-energy X-rays produced by sources located outside the patient. It is the most common form of radiotherapy for treating head and neck cancer. In this thesis only external beam radiotherapy will be considered.

1.1.1 Mechanism of action

Radiotherapy is based on the principle of damaging the deoxyribonucleic acid (DNA) of the malignant cells by the delivery of ionizing radiation. There are various types of DNA lesions that are inflicted either by direct ionization or via charged particles (free radicals) generated in the cells as a result of the irradiation. Cell death is best achieved by breaking the double strand of the DNA. In general, DNA lesions are rapidly repaired by cellular enzymatic pathways. However in some cases, the cells are unable to completely repair the DNA damage leading to a mutation or apoptosis after a variable number of cellular cycles. This mode of cell death is called mitotic and is the major mechanism of tumour response in radiotherapy. Not only does radiation destroy cancer cells, but it also affects dividing cells of normal tissues. Each time a dose of X-ray is delivered, there is a need for balance between destroying cancer cells and minimizing damage to normal cells. Most normal cells are able to recover from the effect of

radiation and repair themselves to proper functioning after some time. For this reason, the total amount of radiation delivered (i.e. dose prescription) is spread out over time (i.e. fractionated). Fractionation gives time for normal cells to recover, while cancer cells are generally less efficient in repair between fractions.

1.1.2 Dose prescriptions

There is no consensus on dose prescription for a given tumour. Indeed, there might be 3 or 4 different recommendations in dose prescriptions with similar variations in fractions but differing in delivery time. For instance, most clinical centers use conventional fractionation with treatment administered on a week day basis for 30 to 40 fractions. In accelerated fractionation, overall total dose is given in a shorter time than in conventional fractionation. This result in greater toxicity and therefore only limited acceleration is possible without altering fraction size. Hyperfractionation refers to the practice of reducing the fraction size of the conventional regime. Treatment is delivered twice or even three times a day in smaller fraction size to enable a higher dose overall to be delivered. Hypofractionation refers to giving a treatment in a shorter time than in the conventional regime, but using bigger doses per day. The total dose administered is also reduced to minimize toxicity.

1.1.3 External radiotherapy

Current radiotherapy is largely based on principles established in the 1940s (Meredith, 1984) when treatments became consistently reported and more quantitative. These principles state that it is necessary to determine the size, shape and position of the volume to be treated, and that this volume should receive a dose distribution as uniform as possible. Conversely, the dose to healthy tissues outside the treatment volume should be minimized, and it is important to give consistent treatments for patients with similar disease type in order to gather information about proper dose levels. External beam radiotherapy tries to meet those principles as closely as possible. With this method, it is impossible to deliver zero dose to healthy tissue through which the beams pass through to the tumour. Therefore, multiple beams are employed from several directions to deliver a cumulative dose to the tumour volume whilst minimising the dose to normal tissue.

In the 1930s, patients were treated with orthovoltage and superficial X-ray units (up to 300 kV). These units deliver high dose to the surface whilst still contributing dose at depth. Cobalt 60 machines were then developed in the 1950s and deliver a higher dose at depth with energy photons in excess of 1 MV. Nowadays, linear accelerators (linacs) are the most common source of high-energy X-ray beams. Modern linacs offer a choice of photon and electron energies. They produce megavoltage photons of 4 to 20 MV in energy which are able to penetrate to the deepest seated tumours in the largest patients. Clinically, 4-8 MV beams are the most useful, providing balance between penetration and adequate surface dose. Figure 1.1 shows the main components of a linear accelerator.

Methods for delivering radiation dose have seen dramatic changes over the past decade. These changes were driven in large by advances in computer technology that led to the development of sophisticated 3D computer-controlled radiation treatment planning systems. The evolution in radiation therapy techniques began with 3D conventional radiation therapy where the shape of the beam was simply square

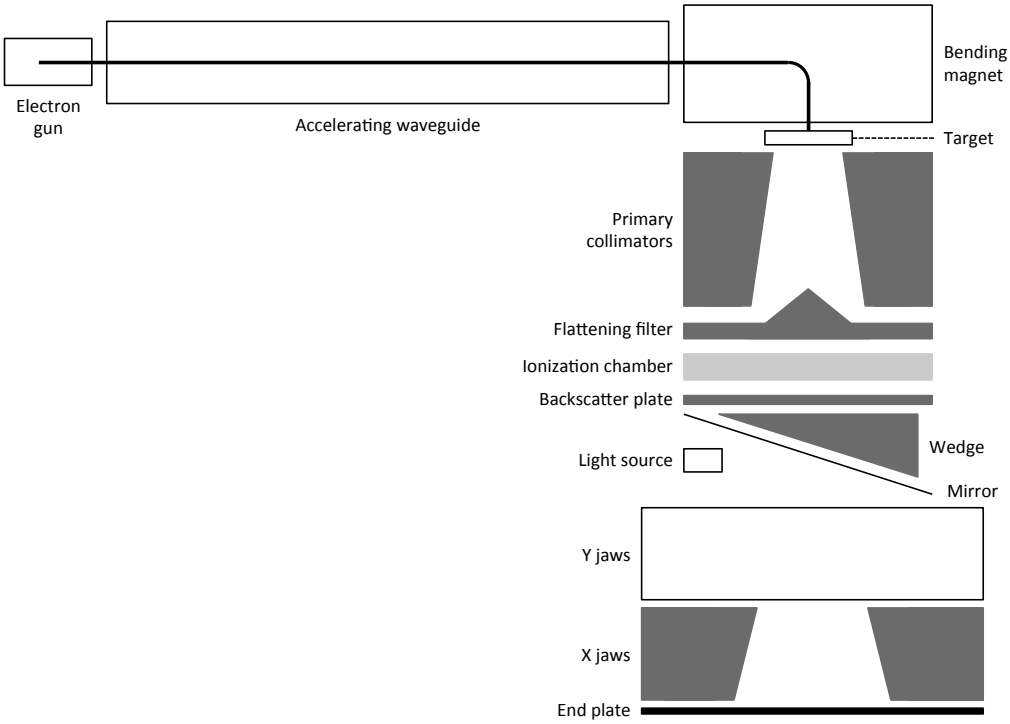

Figure 1.1: Block diagram showing the main components of a linear accelerator[1].

or rectangular. Subsequently, 3D conformal radiation therapy was developed (Dearnaley et al., 1999). In this scheme, the aim of conformality is to design radiation beams that follow the shape of the tumour more closely and conform the spatial dose distribution to the 3D volume to be treated. This allowed better precision of radiation delivery to the volume and improvements in sparing healthy tissues. Another step forward was the development of intensity modulated radiation therapy (IMRT) (Webb, 2001). A uniform dose distribution can be created around the tumour while preventing the surrounding structures from being subject to high doses. This can be achieved by either modulating the intensity of the beam through the linear accelerator using wedges or by use of multi-leaf collimators. Both of these methods alter the fluence of radiation exiting the accelerator. IMRT has now become the standard in treating many cancers, including head and neck cancer. More recently, imaging-guided radiotherapy (IGRT) (Xing et al., 2006) has emerged and enables the tracking of tumour regression and anatomical changes in the surrounding tissue during the whole course of radiotherapy. It is a broad term for radiotherapy techniques which incorporates multi-dimensional imaging modalities into the process of radiotherapy planning. Finally, stereotactic radiotherapy (also called radiosurgery) (Grosu, 2006; Leksell, 1983) is another radiation technique. The radiation delivered has a sharp dose fall-off between the volume to be treated and the surrounding normal tissue, thus allowing very precise delivery of radiation to the tumour while minimizing the radiation dose delivered to the surrounding organs. Stereotactic radiation therapy can be achieved with modified linear accelerators or by using the GammaKnife device (Wu et al., 1990).

[1] source: External Beam Therapy Second Edition. Hoskin, 2012.

1.1.4 Treatment planning

Treatment planning is the process of developing a treatment which produces a dose distribution as uniform as possible to the volume to treat, and as small as possible to surrounding region. Traditionally, it begins with imaging the patient in order to acquire tumour position, size, and shape within the patient and the location of critical structures. Computed tomography (CT) is the most commonly used modality, however magnetic resonance (MR) imaging and positron emission tomography (PET) can also be used as they can provide additional anatomical details such as the extent of the tumour. Various structures are then manually delineated on a series of axial slices to produce 3D volumes. Once the delineation has been obtained, a radiotherapy plan can be designed for each individual patient to meet the treatment goal. Computer models are used to simulate total delivery dose and dose distribution within the patient anatomy, including radiation delivered to normal structures and prescribed dose delivered to tumour cells. This process is called dosimetry and the dose level absorbed by tissue is reported in gray (Gy).

To obtain accurate dose calculations from the patient image, the knowledge of the relative electron or physical density of each voxel of the patient image is needed. This is usually achieved by the use of a look-up table within the treatment planning system (TPS) that converts the CT numbers expressed in Hounsfield Unit (HU) in the CT images to electron or physical density relative to water. TPS are able to correct for density on a voxel-by-voxel basis although correction for large areas of tissue using a bulk density correction (where a single density value is chosen for a whole anatomical region) can also be applied. In low density materials, the radiation will travel further before depositing dose, whereas in high-density materials it will be attenuated more rapidly.

Plans are often assessed with the aid of dose-volume histograms (DVH), allowing the clinician to evaluate the uniformity of the dose applied to the tumour and the sparing of healthy structures. The therapeutic success of a radiation treatment is determined by the balance between tumor control and normal tissue complication probability (refered to as the TCP-NTCP balance). Indeed, in a high proportion of patients the biological dose necessary for tumor eradication can not be delivered because of a high probability of complications due to collateral damage to surrounding tissues.

Finally, planning process can be differentiated between forward and inverse planning. In forward planning, the planner starts the process by choosing the appropriate number and directions for the treatment beams to be used. The planner then goes through an iterative process to alter the available treatment parameters to produce a plan that meets the dose coverage for the tumour and the dose constraint for critical structures. In inverse planning, the planner describe the dose distribution that they want to have at the end of the planning process. This is often described as a series of minimum and maximum dose, mean dose, or dose-volume limit. Computer optimization is then used to develop the most appropriate plan.

1.1.5 Definition of target volumes and organs at risk

The International Commission on Radiation Units and Measurement describes recommendations on how to report treatment volumes in external beam radiotherapy (ICRU, 1999). There are three main volumes to be considered in radiotherapy planning. The first volume is the gross tumour volume (GTV). The GTV

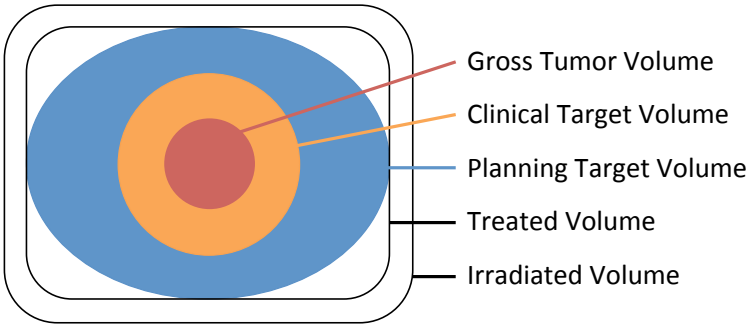

Figure 1.2: Illustration of the main radiotherapy planning volumes.

is essentially the gross demonstrable location and extent of a tumour. It is what can be seen, palpated or imaged. Typically, it is considered that the GTV corresponds to the part of the tumour where the tumour cell density is the highest. Although conceptually the GTV is usually the easiest to define, in practice the edges of the GTV are not necessarily always clear.

The second volume is the clinical target volume (CTV), which contains the GTV plus a margin for sub-clinical disease spread. It is the most difficult volume to define as its definition requires clinical assessment of risk and extent of spread, normally based on historical series rather than the extent of tumour quantified in a specific patient. This volume must be adequately treated if cure is to be achieved.

The third volume is the planning target volume (PTV). It allows for uncertainties in planning or treatment delivery. It is a geometric concept designed to ensure that the radiotherapy dose is actually delivered to the CTV. It is defined to account for all uncertainties in treatment such as organ and patient motions and variations in the position of the GTV or CTV relative to the treatment beam and set up errors. The PTV is function of treatment geometry, because the number of beams and their orientations may impose limitations on the PTV's shape or scope. Figure 1.2 illustrates the 3 different volumes.

Radiotherapy planning must always consider critical normal tissue structures, known as organs at risk (OARs). OARs are normal tissues whose radiation sensitivity influences treatment planning or the prescribed radiation dose. A margin is added to the OAR, which is analogous to the PTV margin around the CTV, and generates the planning organ at risk volume (PRV). It is helpful to create a PRV around an OAR since the loss of normal tissue from radiation damage can result in severe clinical manifestations.

1.1.6 Head and neck cancer

Most HNCs begin in the squamous cells that form the lining of the moist surfaces inside the mouth, the nose, and the throat (Vokes et al., 1993). These squamous cell cancers are often referred to as squamous cell carcinomas of the head and neck. HNCs are categorized by the areas in which they begin: oral cavity, pharynx, larynx, sinuses and nasal cavity, or salivary glands. These areas are labelled in Figure 1.3.

HNCs often spread to the cervical lymph nodes of the neck. This spread is often the first sign of the disease at the time of diagnosis. Lymph node involvement is the most important prognostic factor regarding the survival of patients with head and neck cancer (Snow et al., 1982). Lymph nodes can be

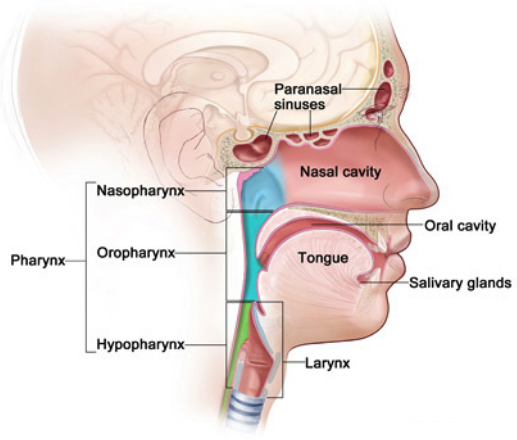

Figure 1.3: Head and neck cancer regions[1].

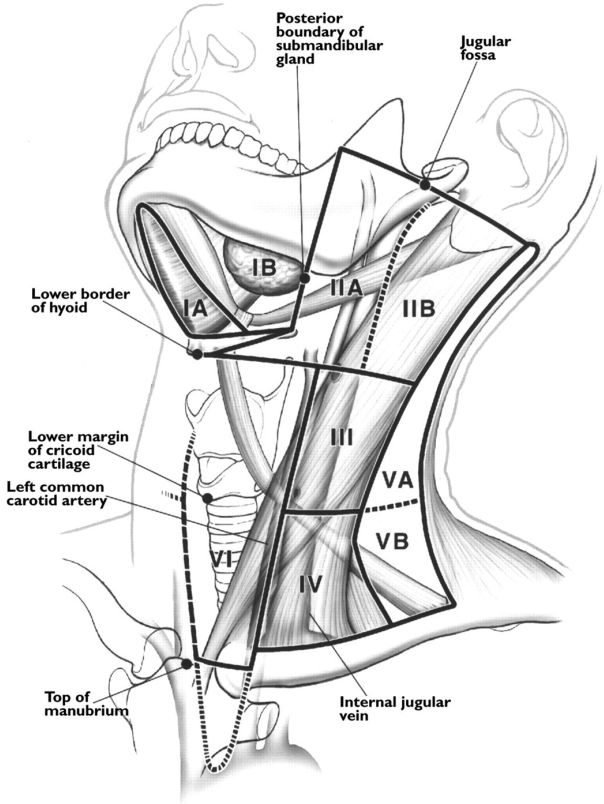

Figure 1.4: Lymph node levels in the neck[2]. IA/IB: submental/submandibular nodes, IIA/IIB: Upper anterior/posterior jugular nodes, III: middle jugular nodes, IV: lower jugular nodes, VA/VB: upper/lower posterior node, VI: central compartment nodes.

classified by level based on anatomic landmarks that can be consistently identified on cross-sectional imaging (Som et al., 2000). Figure 1.4 illustrates this classification .

Determining the extent to which a cancer has spread is mandatory in order to plan effective treatments and remove all tumour tissues. The tumour-node-metastasis (TNM) staging system was developed

[1] source: www.cancer.gov
[2] source: External Beam Therapy. Second Edition. Hoskin, 2012.

Tumour	How much normal tissue tumor has gone into
Tis or (T0)	Carcinoma *in situ*
T1	Tumour $<$ 2 cm
T2	Tumour $>$ 2-4 cm
T3	Tumour $>$ 4 cm
T4	Tumour involves adjacent structures
T4a	Operable disease
T4b	Inoperable disease
Nodes	Spread of cancer to lymph nodes
N0	No regional nodes metastasis
N1	Single ipsilateral node metastasis $<$ 3 cm
N2a	Single ipsilateral node metastasis $>$ 3-6 cm
N2b	Multiple ipsilateral node metastasis $<$ 6 cm
N2c	Bilateral or contralateral $<$ 6 cm
N3	Lymph node metastatis $>$ 6 cm
Metastasis	Spread of tumor beyond lymph nodes to other parts of body
M0	No distant metastasis
M1	Metastasis to distant organs

Table 1.1: Generic staging for head and neck cancer.

to achieve consensus on one globally recognised standard for classifying the extent of spread of a cancer (Edge and Compton, 2010). Generic staging for head and neck cancer is presented in Table 1.1.

1.2 Thesis contributions

The aim of this thesis was to clinically evaluate the feasibility of using atlas-based methods in the context of radiotherapy planning. The contributions of this thesis include the following:

- I propose a new atlas-based segmentation method based on manifold learning. The method is computationally fast and scalable, making it suitable for segmenting large datasets of images acquired during radiotherapy planning. I demonstrate that this method produces segmentation accuracy close to or significantly higher than state-of-the-art methods.

- I demonstrate that atlas-based methods can produce segmentations as comparable as manual contouring in the context of radiotherapy planning and decrease manual labor. Automatic segmentations obtained with my method were graded for clinical acceptability as well as or better than manual contours with a rate of 83%. In addition, I show that overlap measures don't reliably reflect clinical acceptability of a segmentation.

- I demonstrate the feasibility and accuracy of MR imaging-based treatment planning. Synthetic CT images can be generated from MR images using atlas-based methods to support the workflow of radiotherapy planning. My method generated synthetic CT images that showed high similarity with real CT images. The dose distributions calculated on the synthetic CT images were also in good agreement with the original doses used during planning.

1.3 Thesis organization

The next chapter is a review of current methods designed to obtain automatic segmentations using a large dataset of atlases. The concepts of image registration, atlas-based segmentation and feature extraction in a large dataset using manifold learning are presented as well as some of their applications in medical imaging. Inter- and intra-observer variability is an important obstacle in the assessment of automatic

algorithms. Variations in OARs delineation on CT images may unintentionally influence the development and optimization process of those algorithms. In addition, differences in segmentation protocol for radiotherapy planning could be one of the reasons explaining variations in accuracy between algorithms. As a result, to reduce the uncertainty related to the delineation of OARs on CT images, experiments in Chapter 3 are performed on a dataset of MR atlases of the hippocampus, a structure that has been extensively studied in the litterature. MR imaging provides several advantages over CT imaging including improved soft tissue visualisation and hence better target delineation. The inter- and intra-observer variability in the delineation of anatomical structures on MR imaging is reduced compared to delineation on CT images, making the assessment of automatic algorithms more accurate. The findings in Chapter 3 are then applied in Chapter 4 for the segmentation of OARs on CT images for radiotherapy planning. In that chapter, a novel approach for the evaluation of segmentation is proposed and it is demonstrated that atlas-based segmentation can automatically produce clinically acceptable segmentation of OARs, with results as relevant as manual contouring. Chapter 5 demonstrates how atlas-based methods can be used to synthesize an electron density map from MR images. The findings in that chapter show that radiotherapy planning based on MR imaging with synthetic CT images generated through atlas-based method is feasible for head and neck cancer treatment. The doses calculated from synthetic CT images agreed well with those from real CT scans. Finally, Chapter 6 concludes this thesis and outlines some future research directions.

Chapter 2

Atlas-based methods for radiotherapy

Quantitative analysis and volumetric measurements in radiotherapy require the segmentation of anatomical structures. In practice, contouring is often performed by a human operator which is time consuming and labor intensive. It is also subject to inter- and intra-operator variability despite universally accepted delineation guidelines (Fiorino et al., 1998; Scheltens et al., 1997). In addition, the increase in size and availability of imaging databases renders manual segmentation an impractical task to perform, especially for large-scale clinical studies. As a result, much effort has been devoted to developing automatic segmentation methods.

Several automatic segmentation methods have been proposed in the literature such as deformation models (Chupin et al., 2007; Shen et al., 2002), appearance-based models (Duchesne et al., 2002; Hu and Collins, 2007), and atlas-based methods (Aljabar et al., 2009; Rohlfing et al., 2004a). In recent years, atlas-based segmentation methods have been the subject of intensive interest for their accuracy and robustness in segmenting anatomical structures. Those methods make use of image registration and *a priori* anatomical information provided in the form of an atlas. An atlas in this context is a pair of image volumes: one intensity image (referred to as a template) and its associated segmented image (referred to as a label). Atlas-based methods benefit from large dataset to cover the wide range of anatomical variation within a population of images. However with dataset becoming larger and larger, it is crucial to find a compact representation of a population of images. Recently manifold learning methods have been used to model and extract the features of large dataset and used in combination with atlas-based methods to improve their performances. The aim of my thesis is to use atlas-based methods to obtain segmentation of OARs in radiotherapy and to construct a synthetic CT image from a dataset of CT atlases. In this chapter, the concept of image registration, atlas-based method, manifold learning and some of their applications in medical imaging are presented.

2.1 Image registration

An image R can be segmented by establishing spatial correspondences with an atlas in a pairwise anatomically correct way, a process referred to as image registration (Rueckert and Schnabel, 2011). Given an accurate coordinate mapping from R to the atlas, the label for each voxel in R can be determined by looking up the anatomical structure at the corresponding location in the atlas under that mapping. Labelling an image by mapping it to an atlas is known as atlas-based segmentation. Comput-

ing the coordinate mapping between the image and atlas is a critical step in such a method. This section details the fundamental of image registration.

The aim of image registration is to find the optimal geometric transformation which maximizes the correspondence between two images. Image registration can be applied to images from the same subject acquired with different imaging modalities, or at different time points (intra-subject registration). It can also be used to align images obtained from different subjects (inter-subject registration). This process involves 3 main components: a transformation model, an objective function and an optimization method.

2.1.1 Transformation model

The transformation model defines a geometric transformation between the images. There are 3 types of transformation. Each type is characterised by a number of parameters that describes the degree of freedom of the transformation. The first type of transformations is called rigid transformation. It consists of moving an image in space while preserving its original shape. The image can only be translated or rotated. For 3 dimensional images, a rigid transformation is parametrised by 6 degrees of freedom: 3 rotations and 3 translations.

The second type of transformations is called affine transformation. An affine transformation is parametrized by 12 degrees of freedom: 3 translation and 3 rotation parameters as in rigid transformation in addition to 3 scaling and 3 shearing parameters. This transformation is global, meaning that every parameter will affect the whole image. Some affine registration algorithms have been specifically developed for medical imaging (Jenkinson and Smith, 2001; Ourselin et al., 2001).

The third type of transformations model is called non-rigid transformation. In this case, localized transformations are applied to the image. These localized transformations can be defined by a set of displacement vectors (parametric transformations), or by a displacement vector associated with every voxel in the image (non-parametric transformations). As a result, for images containing $256 \times 256 \times 256$ voxels, non-rigid transformations can involve millions of degrees of freedom. Non-rigid parametric transformations developed for medical imaging include the Hierarchical Attribute Matching Mechanism for Elastic Registration algorithm (HAMMER) (Shen and Davatzikos, 2002), and the free-form deformation based on cubic B-splines (Rueckert et al., 1999). Non-parametric transformations include the Demons algorithm (Thirion, 1998), fluid-based algorithms (Christensen et al., 1996; Freeborough and Fox, 1998) and velocity field-based algorithms (Beg et al., 2005). Figure 2.1 illustrates the different classes of transformation.

Once a transformation model $\mathbf{T}$ is chosen, it is used to deform a floating image F into a warped image $F(\mathbf{T})$ that is in the space of a reference image R. For each voxel $\vec{x}$ in R, its transformed position in F is given by $\mathbf{T}(\vec{x})$. The intensity for each voxel in $F(\mathbf{T})$ is then resampled from the original image F. Figure 2.2 illustrates image resampling.

2.1.2 Objective function

An objective function is used to assess the similarity between R and $F(\mathbf{T})$ by mean of a measure of similarity. Various measures have been proposed and they can be classified into feature-based or intensity-based. Feature-based measures require the extraction of points, lines or surfaces and aim to minimize

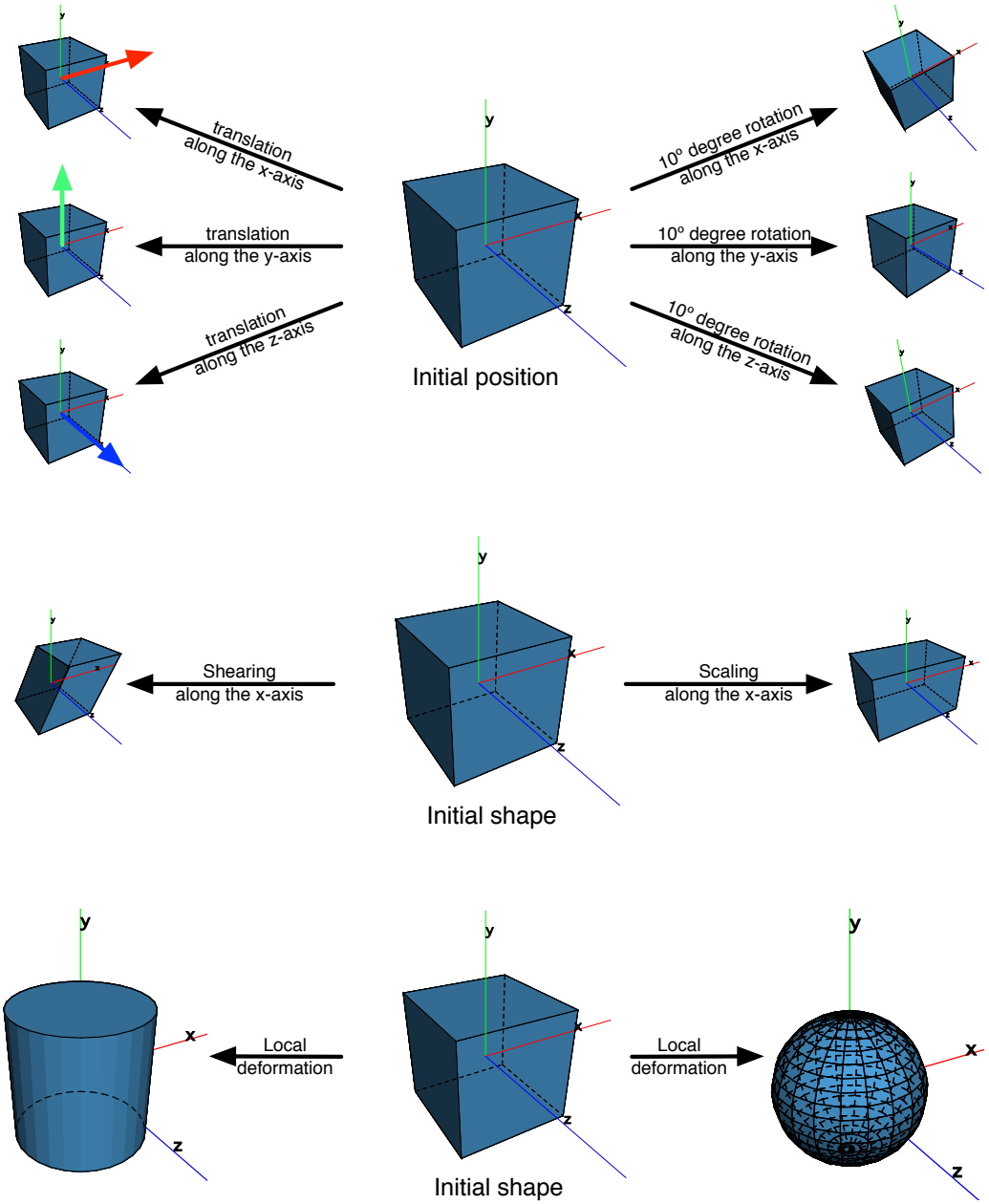

Figure 2.1: Different transformation models applied to a cube[1]. Top: rigid transformation parameters. Middle: affine transformation parameters applied to the x-axis only. Bottom: two non-linear deformations applied to the initial shape.

the distance between the corresponding features in the images. In contrast, intensity-based measure do not require feature extraction and rely on optimizing a voxel-based similarity metric. Popular similarity metrics used in medical imaging are the sum of squared differences (SSD) (Gee et al., 1993) and the normalized cross-correlation (NCC) (Collins et al., 1995; Dong and Boyer, 1995) for mono-modality registrations, and the normalized mutual information (NMI) (Maes et al., 1997; Wells III et al., 1996) for multi-modal registrations. A regularization term might be added to the objective function. This term constrains a transformation model, for example to produce one to one correspondences. A deformation

[1] source: Efficient Dense Non-Rigid Registration using the Free-Form Deformation Framework. Modat, 2012.

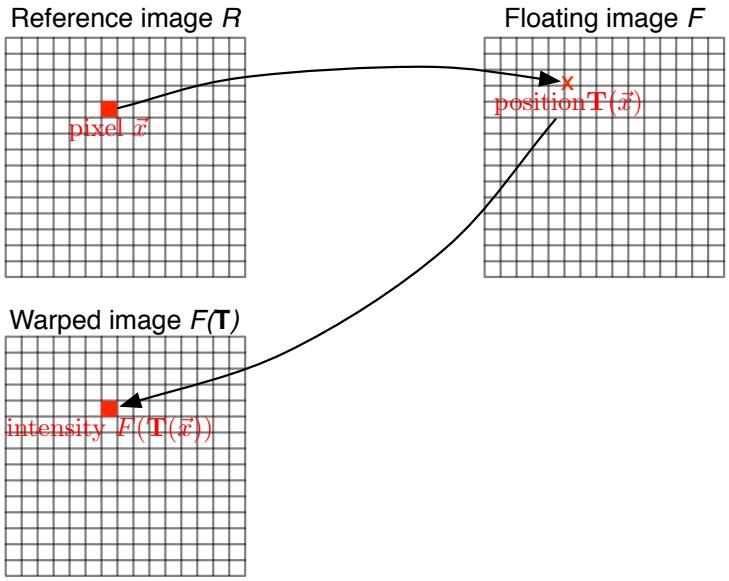

Figure 2.2: Illustration of image resampling[1]. The intensities in the floating image are used to compute the intensities in the warped image.

is deemed realistic when the deformation field is continuous and the topology of the anatomy represented in the image is not broken.

2.1.3 Optimization method

Image registration can be formulated as an optimization problem whose aim is to maximise an associated objective function. Many optimization methods, such as Newton's method (Vercauteren et al., 2009) or the gradient descent (Rueckert et al., 1999), require the estimation of the similarity metric's gradient with respect to the parameters. Other second-order methods may also require an estimate of its Hessian.

This section detailed the fundamentals of image registration. The next section explains how it can be used to obtain the segmentation of an image.

2.2 Atlas-based methods for segmentation

In its simple form, atlas-based segmentation involves performing image registration between a template and an image to be segmented (referred to as a target image). Image registration yields a transformation which allows the label image to be warped and treated as a segmentation estimate for the target. It is then said that the atlas has been propagated to the target. When multiple atlases are propagated to the target, the warped labels need to be combined using a label fusion method to yield a consensus segmentation. The term multi-atlas segmentation is commonly used in the literature (Aljabar et al., 2009; Artaechevarria et al., 2009; Leung et al., 2010) when multiple atlases and a label fusion algorithm are employed for segmentation purposes. Figure 2.3 illustrates the concept of multi-atlas segmentation.

[1] source: Efficient Dense Non-Rigid Registration using the Free-Form Deformation Framework. Modat, 2012.

2.2.1 Atlas propagation

An atlas is defined as a pair of images: a template image A and a corresponding label image L. Let's $\mathbf{T}$ be the transformation that maps the coordinates of the target image R onto those of the template A. For each voxel (x, y, z) in R, the corresponding location in the domain of A is given by $\mathbf{T}(x, y, z)$. The correct label for any voxel (x, y, z) in R can then be calculated through the mapping $L(\mathbf{T}(x, y, z))$.

When a dataset of N atlases $\{(A_1, L_1), (A_2, L_2), ..., (A_N, L_N)\}$ is available, multiple segmentation $L_i(\mathbf{T}(x, y, z))$ of the target can be generated. A label fusion method is then used to combined them into a single segmentation.

Deforming an atlas substantially different from a target requires the usage of non-rigid registration. As a consequence, the potential for registration errors increases with larger deformations. In addition, propagating atlases that are not representative of the anatomical structure of the target or misclassification in the label image will result in an inaccurate segmentation. In the remaining of this section, I detail two fundamental aspects of atlas-based segmentation: atlas selection and label fusion.

2.2.2 Atlas selection

The accuracy of atlas-based segmentation depends on the ability of the image registration to find optimal correspondences between the templates and the target image, which inherently depends on the anatomical similarity between the images. As seen in clinical studies, the range of anatomical variability within

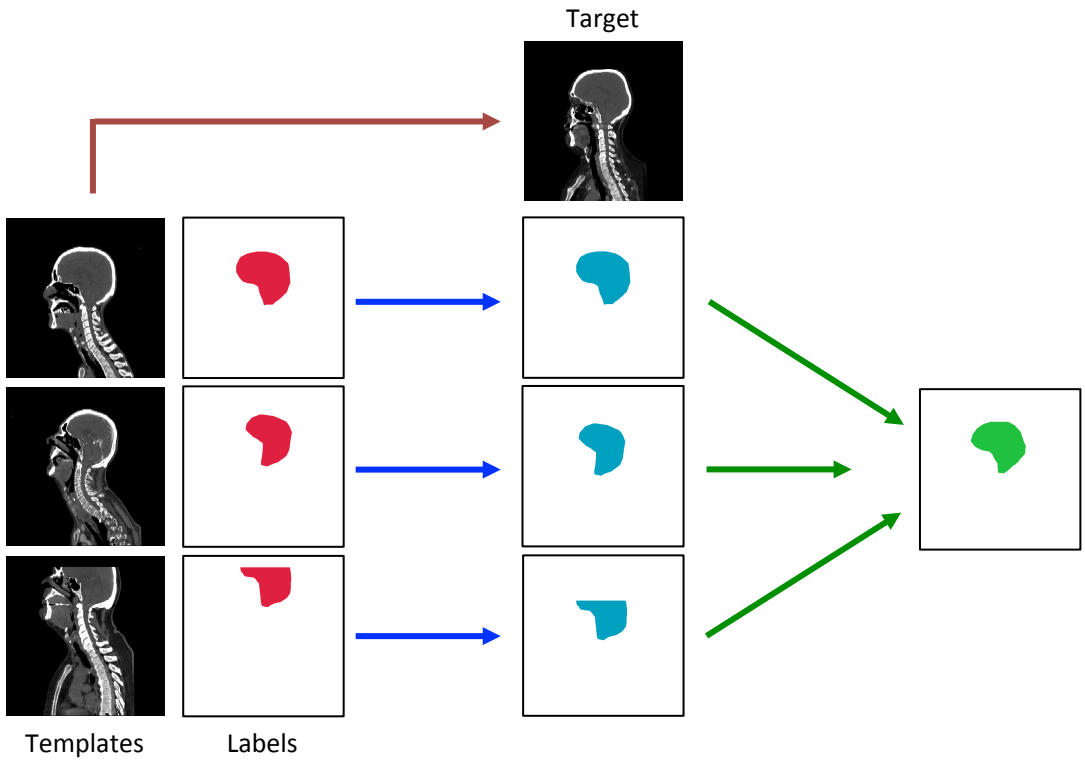

Figure 2.3: Illustration of multi-atlas segmentation. A dataset of templates are registered to a target image (red arrow). The resulting transformations are used to map the corresponding labels onto the target space (blue arrows). The transformed labels are then combined (green arrow) to create an estimate segmentation of the target.

and between subjects can be large. Propagating the atlases that closely match the target reduces registration errors and increases segmentation accuracy. Several atlas selection methods developed for the study of brain structures using MR imaging and radiotherapy using CT imaging are presented in this section.

2.2.2.1 Single atlas selection

In principle, a single atlas can be used to segment a given target image or several different targets (Barnes et al., 2007). This single atlas is often a volume that has been selected from a dataset according to some criteria. In MR based studies, selection has been done in different manners: randomly (Carmichael et al., 2005), based on visual assessment (Rohlfing et al., 2004a), using a standard atlas (Carmichael et al., 2005) such as the MNI 305 atlas (Collins et al., 1994), or based on quantitative analysis such as volumetric measures (Barnes et al., 2007). In Barnes et al. (2008a), the best match from all other subjects in the study was selected based on image similarity in the hippocampal area after affine registration using cross-correlation.

However, an atlas based on a single subject does not represent the wide range of anatomical variation of the human anatomy. For the study of the brain using MR images, several methods have been proposed to build a probabilistic atlas from a set of images to better characterize the variability of anatomical structures within a given population. In this case, information from several images are combined into an average probabilistic atlas using an iterative generation scheme (Brandt et al., 2005; Guimond et al., 2000; Jongen et al., 2004; Rohlfing et al., 2004a). All those methods follow the same framework to construct an unbiased probabilistic atlas. From a given set of images, one is chosen to be a reference image. After all images are registered to that reference, averaging can be performed on voxel intensities (Rohlfing et al., 2004a) or by computing an average transformation (Guimond et al., 2000) to obtain a groupwise image. To obtain a stable atlas, all images are registered to that groupwise image and are subsequently averaged again. The process is repeated until a convergence criteria is reached or after a determined number of iterations. This probabilistic atlas approach has been used for lung segmentation (Li et al., 2003; Sluimer et al., 2005; Zhang et al., 2006a).

Extensive work have been done by Commowick et al. (2006) in developing a single atlas from CT images for radiotherapy planning. In Commowick and Malandain (2007), the atlas in a dataset of images that requires the least amount of deformation was selected for each target image to segment. This selection strategy was based on the fact that small variations between an atlas and a patient to segment improve the quality of registration and the accuracy of the segmentation subsequently. A population-based average atlas was used in Commowick et al. (2008) to segment head and neck anatomical structures. However this resulted in over-segmented structures. A novel framework was introduced in Commowick et al. (2009), where a piece-wise most similar atlas was built from a set of images selected on predefined local regions. This compared favourably to the population-based average atlas. This framework was later improved by Ramus et al. (2010) by combining several selected images for each local region in order to enhance robustness and accuracy. Each image was assigned a set of weights that reflected its similarity to the target within each region.

2.2.2.2 Multiple atlases selection

Propagating multiple atlases and fusing them has recently been shown to be more effective than using a single atlas approach, in particular for segmenting structures in the human brain on MR images (Gousias et al., 2008; Heckemann et al., 2006b; Klein and Hirsch, 2005). By using multiple atlases, the errors due to misclassification or mis-registration can be reduced when the individual labels are fused together. As in the case of a single atlas, multi-atlas methods benefit from selecting atlases similar to the target. Indeed, when using a dataset of atlases that covers a wide range of morphology and pathology, some atlases may be more suitable as candidates for propagation than others. Fusing a large number of atlases with high anatomical structure variability might not yield a valid biological structure. It has been shown that propagating and fusing only suitable atlases produces a better segmentation estimate than using the full dataset (Leung et al., 2010), or a random subset (Aljabar et al., 2009).

Different atlas selection strategies have been proposed in the framework of multi-atlas segmentation in MR based studies. (Aljabar et al., 2009; Klein et al., 2008; Rohlfing et al., 2004a; Wu et al., 2007). A popular approach is to define a similarity metric between the atlases and the target image after registration. This measure is subsequently used to rank the atlases. Segmentation can then be performed using a fixed (Aljabar et al., 2009) or variable (Klein et al., 2008) number of the top-ranked atlases. The similarity metric can be expressed using a variety of metrics, including voxel intensities such as NMI (Aljabar et al., 2009; Klein et al., 2008; Rohlfing et al., 2004a; Wu et al., 2007) or meta-information related to the target (Aljabar et al., 2009). The accuracy achieved by the fusion of different numbers of ranked atlases rises quickly to a maximum and then gradually decline as more and more atlases are added into the fusion process (Aljabar et al., 2009; Leung et al., 2010). This maximum was found to be-tween 7 (Leung et al., 2010) and 25 (Aljabar et al., 2009) depending on the anatomical structure. Leung et al. (2010) developed a multi-atlas propagation and segmentation (MAPS) algorithm for segmenting the hippocampus where the top ranked atlases for a given target were selected based on cross-correlation. Atlas selection can also be done prior to registration by assigning atlases to clusters and registering only some of them (Langerak et al., 2013). These clusters are formed on the basis of the results of pairwise registrations between the atlases. This cluster approach was also used by Blezek and Miller (2007) and Sabuncu et al. (2008). As presented, image similarity has been traditionally used as a direct measure for atlas selection. However, this heuristic criterion might not be able to detect meaningful features for atlas selection within the images.

A framework was presented by Wolz et al. (2010a) in which all the atlases are embedded in a low-dimensional coordinate system using manifold learning to segment to hippocampus on MR images. Only the meaning features of the atlases are encoded in the low-dimensional space. The low-dimensional coordinates provides then a distance metric between images which can be used for atlas selection. The assumption behind using manifold learning is that the intrinsic similarity between images may not be accurately reflected in the high-dimensional space in which the images are represented.

Several studies have employed multi-atlas based segmentation for segmenting anatomical structure in the head and neck region for radiotherapy treatment (Daisne and Blumhofer, 2013; Sjöberg et al.,

2013; Stapleford et al., 2010; Teguh et al., 2011). However, due to the small size of their dataset (5 atlases in Stapleford et al. (2010), 10 in Teguh et al. (2011), 10 in Daisne and Blumhofer (2013), 11 in Sjöberg et al. (2013)), the process of atlas selection was not performed and all atlases were used. Similar to Wolz et al. (2010a), Yang et al. (2010) performs atlas selection within a transformed low-dimensional space, which is more robust to the noises introduced by dental artifacts and registration errors. They also performed atlas selection by visual assessment, where voxel intensities, contrast and head tilt were considered as major factors. The results from the two selection processes were similar. The downside was that the size of their dataset was limited to 10 atlases. Acosta et al. (2011) evaluated different atlas selection strategies for mapping organs (prostate, bladder and rectum) in pelvic CT for prostate cancer radiotherapy planning. The dataset used in their study was significantly larger and included 24 atlases. The cross-correlation (CC), sum of squared differences (SSD) and mutual information (MI) were used to rank a set of atlases according to their similarity with a target image after rigid registration. Results suggested that SSD is a better predictor for mapping than MI and CC. They also found that using the top 20% ranked atlases was a good compromise between accurate segmentation and computational complexity.

2.2.3 Label fusion

When multiple atlases are selected, the choice of a label fusion strategy is required. It is the process of combining multiple label images into a single consensus. It is used to improve segmentation accuracy by averaging out the segmentation errors associated with the mis-registration of some atlases. This process takes place at the voxel level and can be achieved using different strategies. Various fusion methods have been developed specifically for medical imaging.

2.2.3.1 Voting methods

The most widely used fusion method in medical imaging is the majority voting rule (Xu et al., 1992). In this approach, each label image assigns a class at each voxel of the target. The class that received the highest number of agreements is assigned to that voxel.

Individual weights can also be assigned to the label images. Each weight represents the contribution of a given label during the fusion process. The use of weights is based on the assumption that some atlases might be better registered to the target than others and that poor registration will result in inaccurate segmentation. In this case, it is reasonable to give more weight to those well registered atlases during the fusion process. An estimation of the accuracy of the registration between an atlas and a target can be used to assess the influence to be given to that atlas. Weight assignment can be either global (i.e. same weight to every voxel in the image) or local (i.e. one weight per voxel) (Artaechevarria et al., 2009; Isgum et al., 2009; Sdika, 2010). For instance, majority voting is a specific case of weighted atlases in which the weights are global and equal across all label images.

The use of local weights is supported by the fact that global fusion strategy cannot select the locally good regions within different inputs. Local weights take advantage of the fact that registration may be good in some areas while bad in others. An extensive review of voting methods has been done by Artaechevarria et al. (2009). Global and local weights based on normalised cross-correlation, mean

squared difference, mutual information were compared. It was concluded that local methods should be favoured.

However, those methods do not take into account the fact that different atlases may produce similar segmentation errors. Wang et al. (2013) develop a novel voting method that take into account the dependency between atlases and attempts to directly reduce the expected segmentation error in the combined solution. The dependencies were explicitly modelled as the joint probability of two atlases making a classification error at a given voxel.

2.2.3.2 Probabilistic methods

More sophisticated methods than voting for the fusion of the segmentations are also available. Warfield et al. (2004) presented the simultaneous truth and performance level estimation (STAPLE) algorithm for the segmentation of brain images. It has been extensively used in medical imaging studies (Artaechevarria et al., 2009; Klein et al., 2008; Leung et al., 2010). The mathematical framework of this algorithm is presented below.

The STAPLE algorithm

An image to segment is denoted by Y. It contains N voxels and each voxel is denoted Y_i. Let T be an image of size N representing the hidden binary true segmentation of the structure under analysis in Y. Each voxel T_i is assigned a value of 1 if the structure of interest is present, or a value of 0 is the structure is absent at location i. Let D be an $N \times R$ binary matrix describing R candidate segmentations of the structure, obtained either by manual segmentation or an automatic algorithm. This matrix D is similar to T and contains 1 and 0 representing the presence and absence of the structure at each location i. Let $\mathbf{p} = (p_1, p_2, ..., p_R)^T$ and $\mathbf{q} = (q_1, q_2, ..., q_R)^T$ be the sensitivity and specificity of each one of the candidate segmentations R, indexed by j. These parameters p and q are only depend of R and represent a global degree of agreement or disagreement between a candidate segmentation R_j and a segmentation consensus. In order to estimate T, one needs to maximise the log likelihood of the complete data of this problem (D, T) given the set of parameters (p, q). This maximisation can be described as:

$$(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \arg\max_{\mathbf{p},\mathbf{q}} \log f(\mathbf{D}, \mathbf{T}|\mathbf{p}, \mathbf{q}) \tag{2.1}$$

Using the definition of sensitivity and specificity, $\mathbf{p}$ and $\mathbf{q}$ can be described as the "true positive fraction" and "true negative fraction". Thus, p_j and q_j can be represented by:

$$\begin{aligned} p_j &= Pr(D_{ij} = 1|T_i = 1) \\ q_j &= Pr(D_{ij} = 0|T_i = 0) \end{aligned} \tag{2.2}$$

The parameters $p_j, q_j \in [0, 1]$ are assumed to be characteristic of the rater. This model assumes that the candidate segmentations are independent from each other and that the quality of the result of the segmentation is captured by the sensitivity and specificity parameters. Equation 2.1 can then be maximised by an Expectation-Maximization algorithm. The weight variable $w_i^{(k)}$ denotes the expected probability

of the true segmentation at voxel i being equal to 1 at iteration k and is defined as:

$$\begin{aligned} w_i^{(k)} &\equiv f(t_i = 1|\mathbf{d}_i, \mathbf{p}^{(k)}, \mathbf{q}^{(k)}) \\ &= \frac{a_i^{(k)}}{a_i^{(k)} + b_i^{(k)}} \end{aligned} \tag{2.3}$$

with

$$\begin{aligned} a_i^{(k)} &\equiv f(t_i = 1) \prod_j f(d_{ij}|t_i = 1, p_j^{(k)}, q_j^{(k)}) \\ b_i^{(k)} &\equiv f(t_i = 0) \prod_j f(d_{ij}|t_i = 0, p_j^{(k)}, q_j^{(k)}) \end{aligned}$$

and the parameters (p, q) at iteration $(k + 1)$ are optimised by:

$$\begin{aligned} p_j^{(k+1)} &= \frac{\sum_i w_i^{(k)} d_{ij}}{\sum_i w_i^{(k)}} \\ q_j^{(k+1)} &= \frac{\sum_i w_i^{(k)} (1 - d_{ij})}{\sum_i w_i^{(k)}} \end{aligned} \tag{2.4}$$

In the expectation step (Equation 2.3), weights are calculated at each location i and the estimate of the true segmentation is improved based on the sensitivity and specificity of each candidate segmentation. In the maximization step (Equation 2.4), given the new estimate of the true segmentation, the sensitivity and specificity of each candidate segmentation are optimized.

The original STAPLE algorithm was developed for label images containing a single class. A multi-class STAPLE was subsquently proposed by Rohlfing et al. (2004b). Other segmentation methods that build upon the STAPLE algorithm where proposed which take into account the similarity between the atlases and the target. A non-local STAPLE algorithm was developed by Asman and Landman (2012) where atlas intensities were integrated into the expectation process using non-local correspondence model. Cardoso et al. (2013) developed the similarity and truth estimation for propagated segmentations (STEPS) algorithm in which a local ranking strategy for atlas selection based on the locally normalized cross-correlation was added to the STAPLE algorithm.

2.2.4 Evaluation metrics

Overlap measures are often used to quantify the agreement between two segmentations. The Dice similarity coefficient (Dice, 1945) is the most popular measure reported in the literature. It is defined as: $C = 2(|A \cap B|)/(|A| + |B|)$, where $|A|$ (respectively $|B|$) is the number of voxels in the segmented region A (respectively B) and $\cap$ is the intersection between region A and B. The overlap measure can also be reported as a Jaccard index (Jaccard, 1912) defined as: $J = |A \cap B|/|A \cup B|$, where $\cap$ and $\cup$ are the intersection and union between region A and B. Their values range from 0 to 1, where 0 means no overlap, and 1 signifies a perfect match.

Another metric based on distances can also be used. It is referred to as the "Hausdorff distance", d_H, and it measures the maximum distance d of a point in a set X to the nearest point in another set Y. Distances are usually measured from structure boundaries for each axial slice in the image.

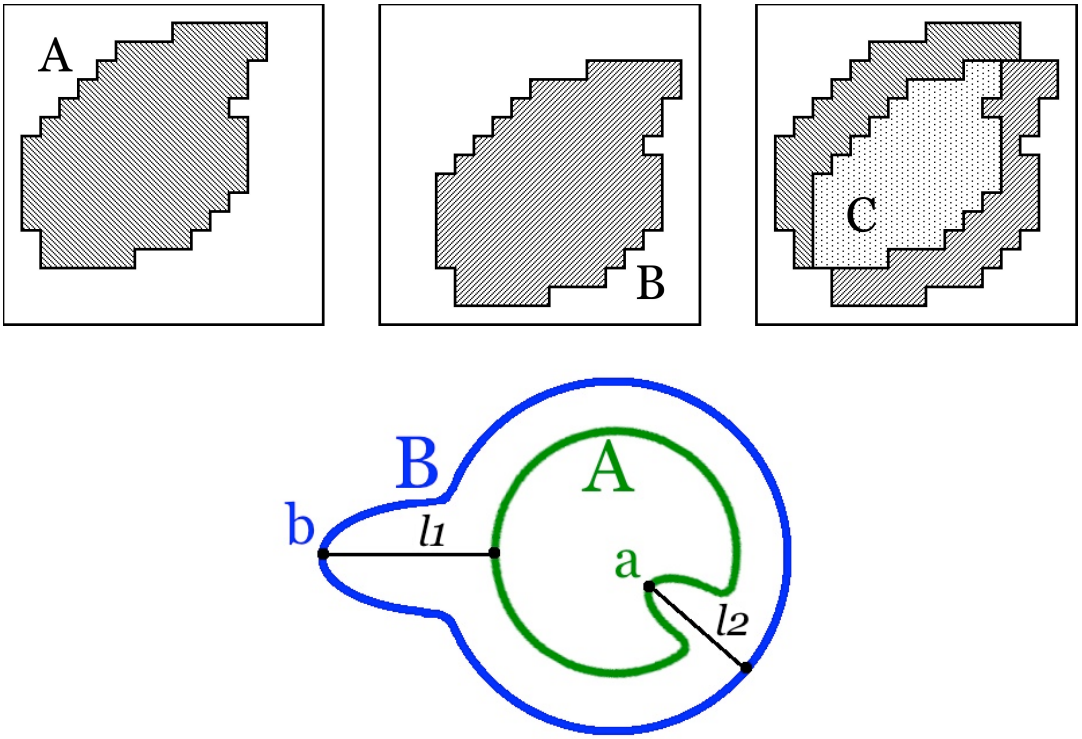

Figure 2.4: Top: illustration of the Dice similarity coefficient. The leftmost picture shows the segmented ROI A, in the middle a different segmentation of the same tissue, ROI B, and in the rightmost picture the two ROIs is put together in the same frame, showing the overlapping area, C. Bottom: schematic figure explaining the concept of Hausdorff distance with two segmentation proposals, *A* and *B*, for a certain structure. The maximum distance from the point *a* on the edge of *A* to edge *B* is marked with the line l_2, and l_1 is the maximum distance from the point *b* on the edge of *B* to *A*. These are the largest minimum distances between the two edges and the Hausdorff distance would in this case be equal to l_1 since $l_1 > l_2$

.

These metrics are quantittive, however they might not reflect the clinical utility of a segmentation. Those different metrics complement each other but are not necessary correlated. Figure 2.4 illustrates the concept of Dice similarity coefficient and Hausdorff distance.

2.3 Atlas-based methods for image synthesis

In recent years, the concept of atlas-based propagation has been used for image synthesis. For instance, it has been used in Burgos et al. (2013) and Marshall et al. (2013) to generate a synthetic CT image to improve attenuation correction for PET/MR scanners. In PET/CT acquisition systems, attenuation maps are derived from CT images. However, in hybrid PET/MR scanners, MR images do not directly provide a patient-specific attenuation map. In Marshall et al. (2013), MR images of patients are compared to a database of CT scans using weighted similarity metrics. Then, a CT scan that closely resembles the patients body type is selected and non-rigidly registered to the MR image. Bones from the registered CT image are then added to the MR image previously segmented into four tissue classes (air, lung, fat and lean tissue) to produce an attenuation map. As presented in the previous section, a set of segmented anatomical atlases from several subjects can be registered to a target image and subsequently fused

according to morphological similarity. This idea was exploited in Burgos et al. (2013) for the propagation and fusion of continuous image intensities. In their method, a dataset of CT/MR paired images from multiple subjects was used to propagate the CT intensities onto an MR target image. A local image similarity measure was used between the target image and the set of registered MR atlases to model the underlying morphological similarity. It was assumed that if two MR images are similar at a certain spatial location, the two corresponding CT images are also similar at this location. This resulted in the synthesis of a patient-specific CT image from which an attenuation map was then generated.

A similar framework was used by Uh et al. (2014) to produce a synthetic CT image for dose calculation in MR imaging based radiotherapy treatment. A dataset of CT/MR atlas image pairs was first constructed by aligning planning CT and MR images of the same patient. The deformation of each CT atlas to the MR target image was then determined by the transformation between the corresponding MR atlas and the MR target image. The multiple deformed CT atlases were subsequently combined to produce a synthetic CT image using a simple arithmetic mean process. This study showed that synthetic CT images generated from multiple deformed atlases are more suitable for treatment planning than those from a single atlas or assigning density values to some segmented areas of the patient volume. The synthetic CT images based showed a high similarity to the real CT images and the corresponding calculated doses agreed well with those based on real CT images.

2.4 Manifold learning

Manifold learning methods have been used to model and extract the features of large dataset and combined with atlas-based methods to improve their performances. Manifold learning has been successfully used in multiple medical imaging applications including segmentation (Zhang et al., 2006b), registration (Hamm et al., 2010; Wachinger and Navab, 2010), classification (Aljabar et al., 2008) and statistical population analysis (Aljabar et al., 2010; Gerber et al., 2010). This section will present the most important manifold learning techniques and how they can be combined with atlas-based segmentation.

2.4.1 Concept

Medical images can be seen as points in a high dimensional space. For instance, a 3D head and neck CT image of size $512 \times 512 \times 256$ voxels may be viewed as vector with more that 67 million dimensions. Each head and neck image has a unique combination of voxel intensities. However, head and neck images share a large degree of anatomical similarity in their appearance. Each individual image may be viewed as a single point in a high dimensional space, but a set of those images may only span over a small region in this space. In other words, medical image sets can be seen as samples of low-dimensional manifolds in the space of all possible images. A number of machine learning techniques called dimensionality reduction techniques have been developed for discovering those manifolds. In recent years, those techniques have been successfully applied in multiple medical applications. In this section, the most common of those algorithms are detailed along with their applications in medical imaging.

2.4.2 Dimensionality reduction

Given a $n \times D$ matrix X where each row represents a set of n vectors (images in this case) $x_i, i \in \{1, 2, ..., n\}$, of dimensionality D, manifold learning aims to discover the intrinsic lower dimensionality d of the dataset X. When using those techniques, images in dataset X are assumed to lie on or near a manifold with dimensionality d embedded in the ambient space of dimension D. d is called the intrinsic dimension of the manifold. Manifold learning transforms a dataset X with dimensionality D into a new dataset Y with dimensionality d (where $d < D$, and often $d << D$), while retaining the geometry of the dataset as much as possible. In the following, a high-dimensional image is denoted by $x_i, i \in \{1, ..., n\}$, where x_i is a i^{th} row of the $n \times D$ matrix X. It is assumed that dataset X is zero-mean.The low-dimensional counterpart of x_i is denoted by y_i, where y_i is the i^{th} row of the $n \times d$ matrix Y. In this section, the most common dimensionality reduction techniques used in medical imaging are presented. They can be classified into linear and non-linear dimensionality reductions techniques. In the literature, the non-linear dimensionality reduction techniques are often referred to manifold learning.

2.4.2.1 Linear methods

Principal Component Analysis

Principal Component Analysis (PCA) (Jolliffe, 2005) is one of the most popular techniques for dimensionality reduction. It constructs a low-dimensional representation of the data that describes as much of the variance in the data as possible using only d principal components. This is achieved by finding the linear mapping function M of size $D \times d$ that maximizes the cost function:

$$trace(M^T cov(X) M) \tag{2.5}$$

where $cov(X)$ is the covariance of the zero-mean matrix X. The principal components of the linear mapping are defined by the first d eigenvectors of the eigenproblem:

$$cov(X)M = \lambda M \tag{2.6}$$

The low-dimensional representation Y of dataset X is then defined as $Y = XM$. The dataset Y can be viewed as a projection of the original data set X onto a new coordinate system constructed by a set of d orthogonal axis (principal components) such that the variance of the data along the first principal component is the greatest. The variance of the data along the second principal component is the second greatest. The subsequent principal components are defined in a similar manner.

Multidimensional scaling

Multidimensional scaling (MDS) (Cox and Cox, 2010) is another classical linear approach that maps the original high-dimensional space to a lower dimensional space by preserving pairwise Euclidean distances. It is based on a pairwise Euclidean distance matrix with elements d_{ij} representing the Euclidean distance between high-dimensional data x_i and x_j. MDS seeks to find the low-dimensional representation that best preserves the pairwise distances in the high-dimensional space. This is achieved by

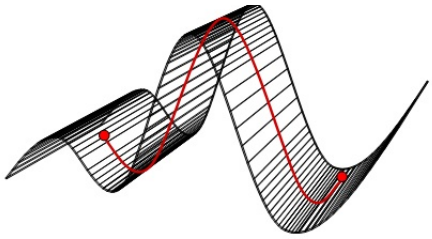

Figure 2.5: Geodesic distance. The geodesic distance between the two red points is the length of the geodesic path, which is the shortest path between the points, that lies on the surface.

minimizing the cost function:

$$\phi(Y) = \sum_{ij} (d_{ij}^2 - \|y_i - y_j\|_{L2}^2) \tag{2.7}$$

where $\|y_i - y_j\|_{L2}$ is the Euclidean distance between two data points in the low-dimensional space. The solution is given by $Y = \lambda^{1/2} V^T$, where V are the eigenvectors of $X^T X$ corresponding to the d eigenvalues, and λ is the top d eigenvalues of $X^T X$. The pairwise distance in MDS does not need to be based on Euclidean distance and can represent various similarity measure between data points.

2.4.2.2 Manifold learning methods

Isomap

In the case where the high-dimensional data lie on or near a curved manifold, MDS might consider two data points as neighbour points, although they might not be on the true underlying manifold. This is because MDS seeks to retain pairwise Euclidean distances only and does not take into account the local distribution of the neighbouring data points. Isomap (Tenenbaum et al., 2000) is a nonlinear embedding technique that tries to solve this problem by preserving pairwise *geodesic* distances between data points as much as possible. The geodesic distance is the shortest path between two points measured over the curved surface of the manifold. Figure 2.5 illustrates the concept of geodesic distance. Isomap estimates the geodesic distances between data points via a neighborhood graph G connecting all n data points. This graph is defined by either connecting every data point x_i to its k nearest neighbours $x_{ij}, j \in \{1, 2, ..., k\}$ or to all data points within some fixed radius ρ. The shortest path between two points in the graph forms an estimate of the geodesic distance between these two points, and can easily be computed using Dijkstra's (Dijkstra, 1959) or Floyd's (Floyd, 1962) shortest-path algorithm. The geodesic distances between all data points X are computed, which result in a pairwise geodesic distance matrix D_G. The low-dimensional representations y_i of data points x_i are computed by applying MDS on the resulting geodesic distance matrix D_G.

Locally linear embedding

Locally linear embedding (LLE) (Roweis and Saul, 2000) is another approach which addresses the problem of high dimensionality by preserving the local properties of the high-dimensional data in the low-dimensional space. The method assumes a locally linear relationship between neighbouring data points. In LLE, each high-dimensional data point x_i is represented as a weighted linear combination of its k

nearest neighbours in the high-dimensional space. This defines a set of weights w_{ij} for the k neighbours of x_i. The aim is to find a low-dimensional representation y_i that respects this weighting. The assumption is that if the low-dimensional data representation preserves the local geometry of the manifold, the reconstruction weights w_{ij} that reconstruct data point x_i from its neighbours in the high-dimensional data representation also reconstruct data point y_i from its neighbours in the low-dimensional data representation. As a consequence, finding the d-dimensional data representation Y with LLE results in minimizing the cost function:

$$\phi(Y) = \sum_{i=1}^{n} \|y_i - \sum_{j=1}^{k} w_{ij} y_{ij}\|_{L2}^2 \tag{2.8}$$

where $\sum_{j=1}^{k} w_{ij} = 1, \forall i \in \{1, ..., n\}$ and subject to $\sum_{i=1}^{n} y_i = 0$ and $Y^T Y = I_d$ with I_d being the $d \times d$ identity matrix. The coordinates of the low-dimensional representations y_i that minimize this cost function are found by computing the eigenvectors corresponding to the smallest d non-zero eigenvalues of $(I_n - W)^T (I_n - W)$, where W is a sparse $n \times n$ matrix whose entries are set to 0 if i and j are not connected in the neighborhood graph, and equal to the corresponding reconstruction weight otherwise and I_n is the $n \times n$ identity matrix.

Laplacian Eigenmaps

A closely related approach to LLE is Laplacian eigenmaps (LEM) (Belkin and Niyogi, 2003) which seeks a low-dimensional data representation by preserving local properties of the manifold. LEM computes a low-dimensional representation of the data in which the distances between a data point and its k nearest neighbours are minimized. In the low-dimensional space, the distance between a data point and its first nearest neighbour contributes more to the cost function than the distance between the data point and its second nearest neighbour. The algorithm first constructs a neighbourhood graph G in which every data point x_i is connected to its k nearest neighbours. Weights w_{ij} are then defined as the similarities between points within a local neighbourhood. For all points x_i and x_j connected in graph G by an edge, the weight of the edge can be computed using a Gaussian kernel function:

$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\rho^2}} \tag{2.9}$$

where ρ is the variance of the Gaussian. When x_i and x_j are not connected in the graph G, w_{ij} is set to zero which result in a sparse adjacency matrix W. The Laplacian eigenmaps embedding is then obtained by minimizing the cost function:

$$\phi(Y) = \sum_{ij} w_{ij} \|y_i - y_j\|_{L2}^2 = 2Y^T L Y \tag{2.10}$$

where $L = D - W$ is the graph Laplacian matrix which is derived from the weight matrix W and the diagonal degree matrix D where $D_{ii} = \sum_{j=1}^{k} w_{ij}$. The Laplacian eigenmaps cost function is optimized under the constraint that $Y^T D Y = I_n$, where I_n is the identity matrix. The low-dimensional data representation Y can then be found by solving the generalized eigenvalue problem:

$$Lv = \lambda D v. \tag{2.11}$$

for the d smallest non-zero eigenvalues. The d eigenvectors v_i corresponding to the smallest non-zero eigenvalues form the low-dimensional data representation Y.

2.4.2.3 Out-of-Sample Extension

An important feature of dimensionality reduction techniques is the ability to embed new high-dimensional data points into an existing low-dimensional data embedding. Indeed, considering a data set X, it would be a computational burden to solve its associated eigenproblem every time new data points are added to it. So-called out-of-sample extensions have been developed for a number of techniques to allow for the embedding of such new data points (Bengio et al., 2004). In PCA, the linear mapping M provides all parameters that are necessary in order to transform new data point from the high-dimensional to the low-dimensional space. In this case, a new high dimensional data point x_{new} and its counterpart point y_{new} in the manifold are related by $y_{new} = x_{new}M$.

For non-linear dimensionality reduction techniques, a direct parametrization of the out-of-sample extension is not available, and therefore, a non-parametric out-of-sample extension is required. Non-parametric out-of-sample extensions perform an estimation of the transformation from the high-dimensional to the low-dimensional space. For Isomap, LLE and LEM, the out-of-sample extension is performed using the Nyström approximation, which approximates the eigenvectors of a large $m \times m$ matrix based on the eigendecomposition of a smaller $n \times n$ submatrix of the large matrix ($n < m$). Experiments on real high-dimensional data have demonstrated the accuracy of out-of-sample extension in positioning an out-of-sample point on a low-dimensional manifold Bengio et al. (2004). For Isomap, the extension is computed with:

$$y_k(x_{new}) = \frac{1}{2\sqrt{\lambda_k}} \sum_{i=1}^{n} v_{k_i}(E_{x_j}[D_x^2(x_i, x_j)] - D_{x_{new}}^2(x_{new}, x_i))$$

where $y_k(x_{new})$ denotes the embedding associated with x_{new}, λ_k and v_k are the k^{th} eigenvalues and eigenvectors of a symmetric matrix derived during the computation of the manifold, $D_{x_{new}}$ denotes the column vector of distances between x_{new} and existing points x, D_x is the matrix of distances between existing points and E_{x_j} represents the column mean of D_x^2.

The extension of LLE is given by:

$$y_k(x_{new}) = \sum_{i=1}^{n} y_k(x_i) w(x_{new}, x_i)$$

where $w(x_{new}, x_i)$ is the weight of point x_i in the reconstruction of x_{new} by its k nearest neighbours from the existing points.

LEM is extended with:

$$y_k(x_{new}) = \frac{1}{\lambda_k} \sum_{i=1}^{n} \frac{1}{n} y_k(x_i) \frac{K(x_{new}, x_i)}{\sqrt{E_{x_i}[K(x_{new}, x_i)] E_{x_j}[K(x_i, x_j)]}}$$

where $K(u, v) = e^{-\frac{\|u-v\|^2}{2\rho^2}}$ is a Gaussian kernel and E is the mean.

2.4.3 Applications in medical imaging

Isomap, Locally linear embedding, and Laplacian eigenmaps are the most commonly used techniques in medical imaging. As an example, Figure 2.6 shows the results of applying different manifold learning algorithms on the same dataset. Some recent findings are presented in the following section.

2.4.3.1 Image registration

Wachinger and Navab (2010) used Laplacian eigenmaps to reduce the computational complexity in multi-modal registration. Aligning images that contain significant intensity variations requires sophisticated similarity metrics such as normalized mutual information that model intensity variation. In their study, Laplacian eigenmaps were used to extract a set of features from MR images which allows the use of the standard L1 or L2 as similarity measures during the registration, effectively decreasing registration time.

When the anatomies of two brain images show large variations in shape and appearance, the registration task turns out to be extremely challenging. Manifold learning was used in Hamm et al. (2010) to tackle the problem of performing large deformation registration. In this case, the underlying manifold of the dataset is approximated by a k nearest neighbours graph based on a pairwise measure between images. Given two images in the k nearest neighbours graph, a transformation between them can be estimated by using the shortest path that connects them on the graph. Each edge in the path represents a transformation between a pair of similar images. An estimated transformation between two images on the graph can be obtained by composing all successive transformations. Since two images connected by an edge on the graph are likely to be very similar, a simple registration algorithm should be enough to accurately register them. Composing those simple registrations should yield a good transformation estimate, even though the two original images are very different. However, the composition of several registrations may lead to the accumulation of small registration errors into larger ones.

2.4.3.2 Image motion parametrization

Manifold learning can also be used to parametrize the transformation between images. Transformation between images corresponds to some form of motion such as cardiac motion or breathing cycle. In this case, temporal ordering is a natural choice for ordering data and a manifold representation may be used to estimate the corresponding image sequence. Isomap was used by Souvenir and Pless (2005) to parametrize cardiac MRI images. In the low dimensional representation, similar points in the cycle were close to each other. This technique aims to characterize the deformations of cardiac MRI image and describe whether images have been deformed due to breathing, or due to contrast agents permeating slowly through the tissues. Wachinger and Navab (2010) took a similar approach to order 4D ultrasound images and place them correctly within the cardiac cycle. Georg et al. (2008) used manifold learning for 4D reconstruction of the lung. In their study, manifold learning enables the estimation of lung volume directly from the images without the need of an external breath measurement. The relation between manifold learning and type of deformations has also been investigated by Souvenir and Pless (2007). It was shown that the pairwise distance measure used for manifold learning should reflect the type of transformation expected. For example, if the images are expected to be related by rigid transformations,

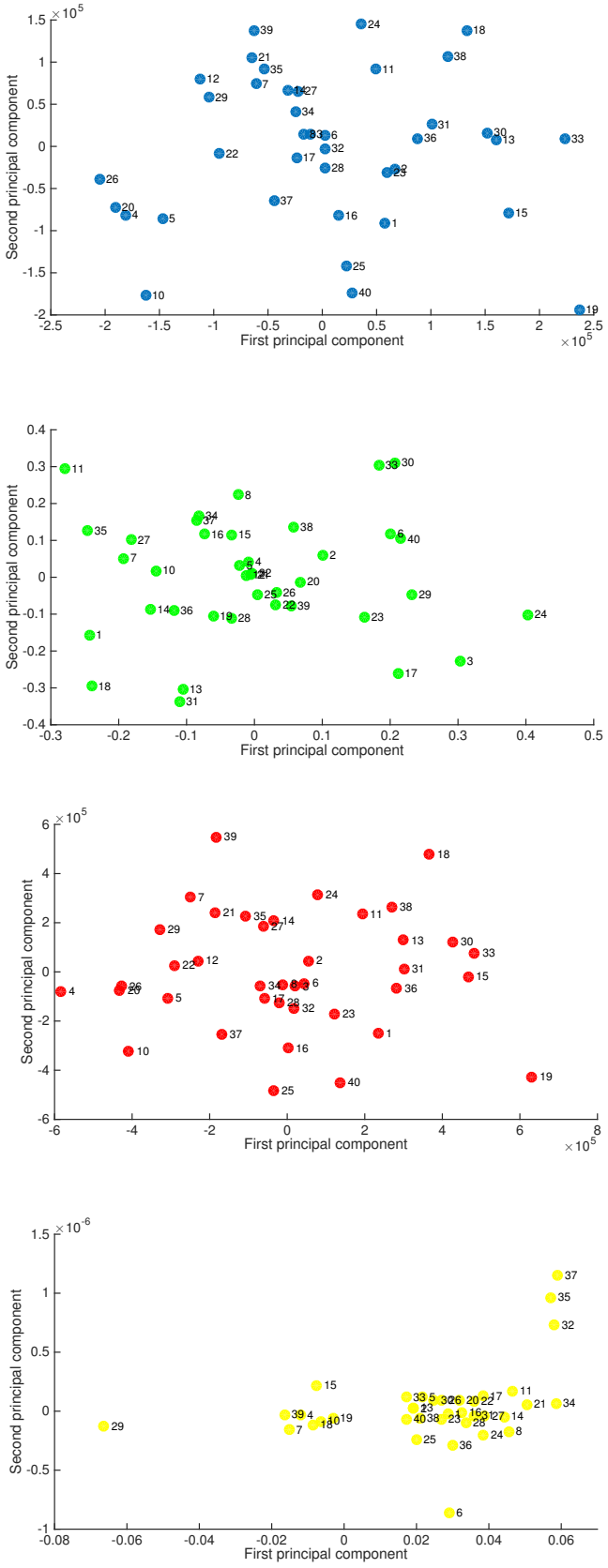

Figure 2.6: Four dimensionality reduction techniques applied to the same dataset. Dataset is composed of 40 head and neck CT images. Only the first and second principal components in the lower dimension are shown. From top to bottom: PCA (blue), Locally linear embedding (green), Isomap (red), Laplacian eigenmaps (yellow). Each number represents and atlas.

then the measure used for manifold learning should be chosen accordingly and may be distinct from one used in non-rigid transformations.

2.4.3.3 Image segmentation

Wolz et al. (2010b) developed the LEAP algorithm (Learning Embeddings for Atlas Propagation) for segmenting a large dataset with a high level of inter-subject variance using a small set of manually labelled atlases that is restricted to a sub-population of the whole dataset. Their method starts with the choice of a pairwise image similarity. This image similarity can be derived from voxel intensities or from distances based on the amount of deformation between images. After computing the low-dimensional representation associated with the dataset, atlases are propagated within the newly defined coordinate system in successive steps. The Euclidean distance in the low-dimensional space is used to define neighbour images. In the first step, the initial atlases are propagated to a number of images in their local neighbourhoods. In the second step, newly segmented images become atlases and are used to subsequently segment images in their vicinity. The process is repeated until the whole dataset is segmented. This approach has the benefit of decreasing registration error as images are only segmented using their most similar counterparts. Results showed that there is a significant improvement in overlap measure between the automated and manual segmentation when applying this manifold-based method in comparison with the direct registration of the available atlas images to all images. Indeed, for very different anatomies between images, large deformations are estimated with a sequence of small deformations which reduces errors resulting from large deformation registration.

2.4.4 Distance metric

An important aspect in manifold learning is the definition of the distance metric used for reconstructing neighbouring points on the manifold. Applications of manifold learning in medical imaging often use a simple distance metric such as the sum of squared differences (SSD) (Georg et al., 2008; Wachinger and Navab, 2010). Metrics developed for the purpose of image registration can also be used as a distance metric in a manifold learning. This includes the normalized cross correlation (Lewis, 1995) or the mutual information (Wells III et al., 1996). Metrics based on voxel intensities measure the similarity of images in terms of their appearance. An alternative approach is to measure similarity based on the shape of the image. For example, deformations produced by image registration can be used to define a distance metric between images. Estimates of local deformation or shape difference instead of measures based on voxel intensity, have been shown to be advantageous for medical image analysis (Pless, 2004; Souvenir and Pless, 2005). A method for building a low-dimensional representation of a set of brain images acquired from AD patients and controls was presented in Gerber et al. (2010). In this case, information about shape variability across the set was captured with a metric derived from non-rigid transformation. Pairwise distances derived from deformation field were also used in Hamm et al. (2010) to implement an efficient large deformation algorithm.

2.5 Summary

In this chapter, the concept of image registration, atlas-based segmentation and manifold learning are presented. A review of the literature has shown that multi-atlas segmentation significantly improves segmentation accuracy compared to segmentation based on a single atlas. It uses a dataset of atlases that is representative of inter-subject variability for a given anatomy. This method has two advantages compared to single atlas propagation. First, it accounts for the anatomical shape variability by using multiple atlases. Second, it is robust because segmentation errors associated with single atlas propagation can be corrected during the fusion process. This method relies on the selection of atlases suitable for propagation as well as the performance of the label fusion algorithm. With the growing amount of imaging data acquired, it is crucial to develop strategies for selecting the best atlases in the framework of atlas-based segmentation in order to achieve optimal accuracy. In recent year, manifold learning have gained popularity in medical imaging. Those algorithm have been used to reduce the complexity inherent to the analysis of medical imaging, such as atlas selection. As presented earlier in Wolz et al. (2010a), manifold learning is used to select atlases which are located in the neighbourhood of the target on the manifold. This novel approach gives promising results on MR images of brains, however it has never been applied on CT images. In addition, each manifold learning technique attempts to preserve a different geometrical property of the underlying manifold. Isomap is a global approach that attempts to preserve pairwise metrics. In contrast, LLE and LEM aim to preserve the local geometry of the data. Since each manifold learning technique is associated with a different objective function, it is legitimate to assume that, for a given data set, the associated embeddings are also different.

The next chapter, Chapter 3, I investigate the appropriate choice of manifold learning technique and manifold parameters that result in optimal atlas selection and subsequently achieve optimal segmentation accuracy. This investigation is done on a dataset of MR images of the hippocampus, a well studied structure in the literature. The results from this investigation are then used in Chapter 4 for segmenting OARs on CT images of head and neck for radiotherapy. In order to study the effect of manifold learning, it was best to experiment on a small and well defined structure such as the hippocampus on a high contrast modality than on full CT images of head and neck which can show tremendous variations due to field of view or patience corpulence.

Chapter 3

Atlas selection using manifold learning

3.1 Introduction

As presented in the previous chapter, multi-atlas segmentation relies on the selection of atlases that are best mapped to a new target image after registration and manifold learning has been proposed as a method for atlas selection. Each manifold learning technique seeks to optimize a unique objective function. Therefore, different techniques produce different embeddings even when applied to the same data set. Previous studies used a single technique in their method and gave no reason for the choice of the manifold learning technique employed nor the theoretical grounds for the choice of the manifold parameters. In this chapter, I compare the results given by 3 manifold learning techniques (Isomap, Laplacian Eigenmaps and Locally Linear Embedding) side-by-side on the same data set. The ability of those 3 different techniques to select the best atlases to combine in the framework of multi-atlas segmentation is assessed. First, a leave-one-out experiment is used to optimize the proposed method on a set of 110 manually segmented atlases of hippocampi and find the manifold learning technique and associated manifold parameters that give the best segmentation accuracy. Then, the optimal parameters are used to automatically segment 30 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI). For this dataset, the selection of atlases with Locally Linear Embedding gives the best results. The findings in this chapter show that selection of atlases with manifold learning leads to segmentation accuracy close to or significantly higher than the state-of-the-art method and that accuracy can be increased by fine tuning the manifold learning process. Those findings are then used in Chapter 4 to segment OARs on CT images of the head and neck.

3.2 Related publications

- **Hoang Duc A.K.**, Modat M., Leung K.K., Cardoso M.J., Barnes J., Kadir T., and Ourselin S. for The Alzheimers Disease Neuroimaging Initiative. Using Manifold Learning for Atlas Selection in Multi-Atlas Segmentation. (2013). PLoS one 8(8): e70059.

- **Hoang Duc A.K.**, Modat M., Leung K.K., Kadir T., and Ourselin S.: Manifold Learning for Atlas Selection in Multi-Atlas Based Segmentation of Hippocampus. (2012). SPIE.

3.3 Methods

3.3.1 Overview

This study aims to qualitatively and quantitatively assess the selection of atlases to combine in the framework of multi-atlas segmentation using 3 different manifold learning techniques. I consider Isomap (Tenenbaum et al., 2000), Locally Linear Embedding (LLE) (Roweis and Saul, 2000) and Laplacian Eigenmaps (LEM) (Belkin and Niyogi, 2003) since those techniques are the most widely used in medical imaging.

My method can be summarized in 3 steps. First, a low-dimensional manifold is learned from the space spanned by the set of atlases using the 3 different techniques (§3.3.2). The neighbourhood relationship on the manifold is derived from non-rigid transformations that align atlases to each other in the high-dimensional space (§3.3.3). Second, a new target image is embedded onto the previously computed manifold by means of the out-of-sample extension (Bengio et al., 2004) (§3.3.4). Third, the target image is segmented using atlases that are within its vicinity on the manifold (§3.3.5).

For each manifold learning technique, I investigate the effects of (i) the number of dimensions of the resulting embedding, (ii) the number of neighbours used to build the k-nearest neighbour graph in the high-dimensional space, and (iii) the number of atlases used during the combination process.

An atlas data set composed of 110 manually segmented images of hippocampi from the MIRIAD public data set (www.ucl.ac.uk/drc/research/miriad) is used to optimize each manifold learning technique on a leave-one-out experiment (§3.4.1). Segmentation accuracy is then validated on an independent set of 30 manually segmented images from the Alzheimer's Disease Neuroimaging Initiative (ADNI, www.loni.ucla.edu/ADNI/) (§3.4.2). The MIRIAD data set is described in §3.3.6. The ADNI data set is described in §3.3.7.

3.3.2 Manifold learning

Given a set of n atlases $A = (a_i, ..., a_n) \in \mathbb{R}^D$, the goal is to identify atlases that are most similar to a target image $x \in \mathbb{R}^D$ using manifold learning. It has been suggested that the set of brain images has an intrinsic dimensionality meaning that points in data set A and image x are lying on or near a manifold with dimensionality d which is embedded in the D-dimensional space (Gerber et al., 2010). By using manifold learning, data set $A \in \mathbb{R}^D$ is transformed into a new dataset $Y = (y_1, ..., y_n) \in \mathbb{R}^d$ with $d << D$, while preserving the non-linear geometry and neighbourhood information of the high-dimensional data in the low-dimensional space. The atlases that are nearest to x are identified on the low-dimensional manifold and used for segmentation.

Variation in brain images is best described by non-linear dimensionality reduction models compared to linear ones like Principal Component Analysis (PCA) or Multi-Dimensional Scaling (MDS) (Gerber et al., 2010). In this study, low-dimensional embeddings are computed with 3 different non-linear techniques: Isomap (Tenenbaum et al., 2000), Locally Linear Embedding (LLE) (Roweis and Saul, 2000) and Laplacian Eigenmaps (LEM) (Belkin and Niyogi, 2003). The differences between those

3 techniques are emphasized by their unique objective functions. For Isomap, the objective function is:

$$\phi(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}^2 - \|y_i - y_j\|_{L2}^2) \tag{3.1}$$

where d_{ij} represents the geodesic distance between a_i and a_j in the high-dimensional space. For LLE, the objective function is:

$$\phi(Y) = \sum_{i=1}^{n}\|y_i - \sum_{j\in N_k(i)} w_{ij}y_{ij}\|_{L2}^2 \tag{3.2}$$

where $N_k(i)$ are the k-nearest neighbours of a_i and weight w_{ij} is the contribution of a_j in reconstructing a_i in the high-dimensional space. As demonstrated by Roweis and Saul (2000), the optimal weights w_{ij} are obtained through minimization by solving a least-squares problem.

Finally, the objective function associated with LEM is:

$$\phi(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\|y_i - y_j\|_{L2}^2 \tag{3.3}$$

where $w_{ij} = e^{-\|a_i - a_j\|^2/2\sigma^2}$ is a Gaussian kernel. All 3 techniques require the construction of a connected graph in the high-dimensional space using the k-nearest neighbour algorithm. The number of neighbours used to build this connected graph is defined as k_D.

Unlike PCA, the embedding produced by these techniques is a function of a metric which determines the k_D-nearest neighbours in the high-dimensional space and subsequently the neighbouring images on the low-dimensional manifold. The metric presented in §3.3.3 is used to find those k_D-nearest neighbours.

3.3.3 Distance between pairs of images

I derive the metric from the method presented by Commowick and Malandain (2007). An atlas a and target image x are similar when the non-rigid transformation that aligns them produces a small deformation. Similarity is based on the displacement field $F_{x\to a}$ of the non-rigid transformation $T_{x\to a}$. In order to avoid the computational load of performing registrations between all atlases and every new unseen target image, an average atlas M is built from the atlases in the data set using the iterative groupwise registration scheme described by Rohlfing et al. (2004a). This enables M to lie near the center of the space of all atlases. From the average atlas M, a displacement field $F_{M\to a}$ (resp. $F_{M\to x}$) is derived from the non-rigid transformation $T_{M\to a}$ (resp. $T_{M\to x}$) for each atlas a (resp. new target x). The similarity is then evaluated with:

$$s(x,a) = \sum_{l=1}^{V}\|F_{M\to a}(l) - F_{M\to x}(l)\|_2 \tag{3.4}$$

where $\|.\|_2$ is the L2 norm and V is the number of voxels in each atlas.

In this framework, the similarity between x and any atlases a can be evaluated by registering x to M. Since M lies near the center of the space of all atlases, the manifold resulting from the approximation

of $F_{x\to a}$ with $F_{M\to a} - F_{M\to x}$ minimizes the error in estimating the neighbourhood relationship when compared to the manifold resulting from the direct computation of $F_{x\to a}$.

The non-rigid transformation T is performed using an efficient implementation (Modat et al., 2010) of the free-form deformation algorithm (Rueckert et al., 1999). The transformation model is parameterized using a cubic B-Spline scheme and the transformation T is driven by the normalised mutual information.

3.3.4 Extending a manifold with a new target image x

For Isomap, LLE and LEM, the out-of-sample extension is performed using the Nyström approximation (Bengio et al., 2004). Experiments on real high-dimensional data have demonstrated the accuracy of out-of-sample extension in positioning an out-of-sample point on a low-dimensional manifold (Bengio et al., 2004). The metric presented in §3.3.3 is also used for extending the manifold.

Since the low-dimensional manifold is embedded in a Euclidean space, the L2 distance is used to determine the k_d-nearest neighbours of x on the manifold. Those k_d-nearest neighbours are subsequently used for label propagation.

3.3.5 Segmentation by fusion strategy

STAPLE (Warfield et al., 2004) is used to combine multiple segmentations generated from the most similar atlases. In a previous study (Leung et al., 2010), it was found that STAPLE gives better results compared to a voting rule or shape-based averaging method when using the MIRIAD data set. It simultaneously computes a probabilistic estimate of the true segmentation and a measure of the performance level (sensitivity and specificity) represented by each segmentation in an expectation-maximization framework. An iterative Markov random field optimized with mean field approximation is used to provide spatial consistency in the probabilistic estimate of neighbouring voxels. The STAPLE algorithm is solved only in the non-consensus area in order to reduce bias as suggested by Rohlfing et al. (2004a). I denote by k_d the number of atlases used for label propagation.

3.3.6 Atlas data set of 110 hippocampi

The MIRIAD data set is used as the atlas data set. It is a database of volumetric MRI brain scans of patients suffering from Alzheimer's disease and healthy elderly people. The data set is publicly available (www.ucl.ac.uk/drc/research/miriad) in anonymised form to aid researchers in developing new techniques for the analysis of serially acquired MRI. The atlas data set consists of 55 subjects who were recruited from the Cognitive Disorders Clinic at The National Hospital for Neurology and Neurosurgery, into a longitudinal neuroimaging study. All subjects underwent clinical assessment including the Mini-Mental State Examination (MMSE) (Folstein et al.,1975). All subjects gave written informed consent to take part in this study. Imaging data were used to create an average atlas using the groupwise registration algorithm described in §3.3.3 and in the parameter optimization process in §3.4.1. Subjects included 36 clinically diagnosed probable AD patients and 19 age-matched healthy controls. All patients fulfilled standard NINCDS/ADRDA criteria (McKhann et al., 1984) for the diagnosis of probable AD. Subject demographics can be seen in Table 3.1. T1-weighted volumetric MR brain scans were performed on the

same 1.5-T Signa unit (General Electric, Milwaukee), using an inversion recovery prepared fast SPGR sequence and a 256×256 image matrix with the field of view being 18 cm (acquisition parameters: repetition time = 15 ms; echo time = 5.4 ms; flip angle = 15°; inversion time = 650 ms). The volumetric scans were reconstructed as 124 contiguous 1.5-mm coronal images. T1-weighted volumetric scans were evaluated by one rater. All scans were N3 corrected (Sled et al., 1998) and bias correction was performed.

Table 3.1: Subject demographics in control and probable AD subjects used for parameter optimization. Mean (SD) unless specified otherwise.

	Control (n=19)	AD (n=36)
Age, years	68.7 (7.0)	69.6 (7.3)
Gender male (%)	9 (47%)	14(39%)
MMSE at baseline, /30	29.5 (0.7)	19.4 (4.1)

The left and right hippocampal regions were manually segmented by an expert segmentor S. The intra-rater variability measured by an ICC is 0.98, based on same-scan analysis of 20 subjects segmented twice. The hippocampus was always measured on the right-hand side of the presented image with the expert segmentor blinded to the subjects name, diagnosis, and left- right orientation of the scans. Each hippocampus took approximately 45 min to delineate (1.5 h per scan). The left hippocampal segmentations from all 55 subjects are flipped along the mid-sagittal plane. This flipping effectively doubles the size of the data set by allowing, for example, the left hippocampus of a target image to be matched to the right hippocampus in the atlas data set. Therefore, the final atlas data set consists of 110 hippocampal images.

3.3.7 ADNI data set of 30 subjects

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.adni.loni.ucla.edu). The 30 ADNI subjects (10 AD, 10 MCI and 10 controls) used for method validation consist of preprocessed baseline volumetric T1-weighted MR images acquired using 1.5T scanners (GE Healthcare, Philips Medical Systems or Siemens Medical Solutions) at multiple sites from the ADNI website. Representative imaging parameters were TR = 2400 ms, TI = 1000 ms, TE = 3.5 ms, flip angle = 8°, field of view = 240×240 mm and 160 sagittal 1.2 mm-thick slices and a 192×192 matrix yielding a voxel resolution of 1.25×1.25×1.2 mm^3, or 180 sagittal 1.2 mm-thick slices with a 256×256 matrix yielding a voxel resolution of 0.94×0.94×1.2 mm^3. The details of the ADNI MR imaging protocol are described in Jack et al. (2008), and listed on the ADNI website (www.loni.ucla.edu/ADNI/Research/Cores/). Each scan underwent a quality control evaluation at the Mayo Clinic (Rochester, MN, USA). Quality control included inspection of each incoming image file for protocol compliance, clinically significant medical abnormalities, and image quality. The T1-weighted volumetric scans that passed the quality control were processed using the standard ADNI image processing pipeline, which included post-acquisition correction of gradient warping (Jovicich et al., 2006), B1 non-uniformity correction (Narayana et al., 1988) depending on the scanner and coil type, intensity non-uniformity correction (Sled et al., 1998) and phantom based scaling correction (Gunter et al.,

2006) with the geometric phantom scan having been acquired with each patient scan.

Table 3.2 shows the clinical and demographic data of the 30 ADNI subjects. The same expert segmentor S as previously mentioned manually delineated the left hippocampus of those subjects. A segmentor S2 also manually delineated the left hippocampus on the same baseline images. The inter- and intra-rater reliability correspond to a Dice's similarity index of 0.93 and 0.96 respectively.

Table 3.2: Subject demographics in set of 30 labelled randomly selected subjects used for method validation. Mean (SD) unless specified otherwise.

	Control (n=10)	MCI (n=10)	AD (n=10)
Age, years	78.6 (5.4)	75.3 (8.8)	77.2 (6.8)
Gender male (%)	6 (60%)	7 (70%)	7 (70%)
MMSE, /30	29.5 (0.7)	27.4 (1.8)	27.0 (2.7)

3.4 Experiments

3.4.1 Optimizing manifold learning parameters using data set of 110 atlases

A leave-one-out approach that excludes both the left and right hippocampi of the target image from the library of 110 atlases is used to optimize the parameters for each manifold learning technique. The following 4-step procedure is repeated for each atlas a_{out} in the library. (i) After excluding a_{out} and its flipped image from the library, an average atlas M is built from the remaining 108 images in the data set. Distances between remaining atlases are computed based on the non-rigid transformations that align them to M as described in §3.3.3. (ii) A manifold is computed from the remaining 108 atlases. (iii) The embedding is extended with a_{out}. Distances between a_{out} and the remaining atlases are derived by registering it to M and performing subtraction of displacement fields. (iv) Its k_d-nearest neighbours are identified on the manifold using the L2 norm and combined in STAPLE to yield an estimated segmentation of a_{out}.

Dice's similarity index (Dice, 1945) is used for evaluation and is computed by measuring the overlap between the estimated segmentation and the manual segmentation. Dice's similarity index is defined as $DS(A,B) = 2|A \cap B|/(|A| + |B|)$, where A is the set of voxels in the automated region and B is the set of voxels in the manual region. A Dice's similarity index is calculated for each a_{out} and a mean Dice's similarity index $\overline{DS}$ is calculated by averaging all 110 scores.

There is no defined procedure to establish the number of dimensions d of a learned manifold, and the number of neighbours k_D to build the connected graph in the high-dimensional space is often determined empirically. Results are evaluated for 3 different techniques: Isomap, LLE and LEM with dimension $d \in [1, 25]$ and a neighbourhood number of $k_D \in [3, 25]$ for each manifold technique. Using STAPLE with a MRF strength of 0.2, segmentations are generated by combining the closest $k_d \in [1, 25]$ neighbours to a_{out} in the lower dimensional space. For LEM, σ is set to 1. A 4D matrix of mean Dice's similarity indexes is then computed with the following axes: manifold type $\in \{ISO, LLE, LEM\}$, $d \in [1, 25]$, $k_D \in [3, 25]$, and $k_d \in [1, 25]$. The coordinates in this matrix that give the highest $\overline{DS}$ indicate the best manifold learning technique with optimized parameters for this data set.

In order to compare atlas selection *with* manifold learning to atlas selection *without* manifold learning, I also compute the results given by: a) a plain k_d-nearest neighbour selection in the high-dimensional space D and b) a k_d-nearest neighbour selection after performing a Principal Component Analysis. I refer to those 2 selection methods as BASE and PCA and results are computed for $k_d \in [1, 25]$. In the BASE method, for each a_{out}, its k_d-nearest neighbours are computed using the metric defined in §3.3.3 and combined in STAPLE to yield an estimated segmentation. As before, a Dice's similarity index is calculated for each a_{out} and a mean Dice's similarity index $\overline{DS}$ is calculated by averaging all 110 scores.

3.4.2 Method validation using data set of 30 ADNI subjects

For method validation, the left hippocampus in the baseline images of 30 randomly selected subjects in the ADNI database (10 AD, 10 MCI and 10 controls) were segmented. Those images differ from the MIRIAD data set of atlases used for parameter optimization. The atlas data set of 110 images is used to segment each of the ADNI target images. The optimal parameters determined in §3.4.1 are used to generate left hippocampal regions. Since the right hippocampus segmentations for this set of 30 subjects were not available, I only evaluate the accuracy of my method on the left hippocampus.

3.5 Results

3.5.1 Results from method optimization

The best combination of manifold learning technique and parameters is Locally Linear Embedding with a manifold dimension of $d = 11$, a neighbourhood size $k_D = 23$ and combining the top $k_d = 7$ matches in STAPLE, giving a mean (SD) Dice's similarity index $\overline{DS_{max}}$ of 0.9077 (0.0211). In contrast, Isomap and Laplacian Eigenmaps resulted in Dice's similarity indexes of 0.8995 (0.0228) and 0.8971 (0.0245) with $d = 21$, $k_D = 23$ and $k_d = 9$ and $d = 13$, $k_D = 21$ and $k_d = 19$ respectively. Each graph in Figure 3.1 shows the mean Dice's similarity index for each manifold learning technique when d, k_D and k_d are fixed to their respective optimal parameters. It is interesting to note that all 3 manifold learning techniques result in a very high mean Dice's similarity index (>0.89). Using a 2-tailed paired t-test, Locally Linear Embedding gives a significantly ($p = 0.0216 < 0.05$ and $p = 0.0275 < 0.05$) higher average Dice's similarity index compared to Isomap and Laplacian Eigenmaps, whereas the difference between Isomap and Laplacian Eigenmaps is not statistically significant ($p = 0.3250 > 0.05$). The accuracy achieved by fusing multiple segmentations quickly rises to a maximum and then gradually declines as the number of segmentations increases. This is in line with results published in Aljabar et al. (2009) and Leung et al. (2010): the gradual decline corresponds to adding dissimilar images into the combination process, resulting in segmentation errors. The accuracy also flattens out for manifolds of 3 or more dimensions. This suggests that this data set of hippocampi can be described mostly by 3 main modes of variation, and this is consistent across all manifold learning techniques presented. The number of neighbours k_D used to build the connected graph has little effect on the accuracy when using Isomap and Laplacian Eigenmaps. In contrast, increasing k_d increases the accuracy achieved with Locally Linear Embedding.

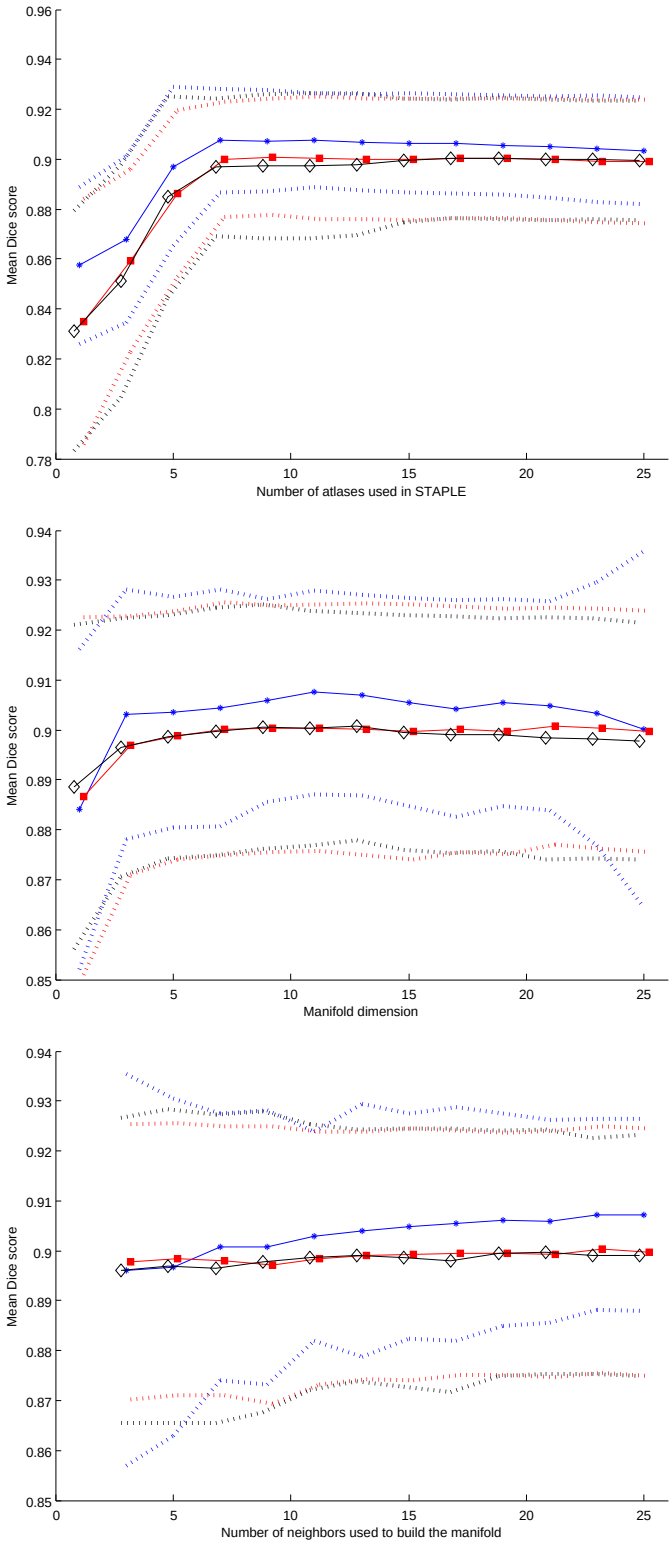

Figure 3.1: Mean Dice's similarity index computed for $k_D \in [3, 25]$, $d \in [1, 25]$, $k_d \in [1, 25]$. Locally Linear Embedding is in blue, Isomap is in red and Laplacian Eigenmaps is in black. Solid lines represent the mean Dice's similarity index, doted lines represents the standard deviation. Mean Dice's similarity index against: (a) the number of atlases fused in STAPLE (d and k_D fixed to best parameters), (b) the neighbourhood size k_D in computing the manifold (d and k_d fixed to best parameters), and (c) the manifold dimension d (k_D and k_d fixed to best parameters).

Table 3.3: Mean Dice's similarity indexes $\overline{DS}$ (SD) obtained with manifold learning selection (LLE, ISO, LEM) and BASE/PCA methods. p-values comparing each approach with each other are reported.

	LLE $d = 11$, $k_D = 23$, $k_d = 7$	ISO $d = 21$, $k_D = 23$, $k_d = 9$	LEM $d = 13$, $k_D = 21$, $k_d = 19$
Mean DS (SD)	0.9077 (0.0211)	0.8995 (0.0228)	0.8971 (0.0245)
p-value	LLE vs. ISO, $p = 0.0216$ LEM, $p = 0.0275$ BASE, $p = 0.0056$	ISO vs. LLE, $p = 0.0216$ LEM, $p = 0.3250$ BASE, $p = 0.0137$	LEM vs. LLE, $p = 0.0275$ ISO, $p = 0.3250$ BASE, $p = 0.0204$

	BASE $k_d = 9$	PCA $k_d = 11$
Mean DS (SD)	0.8756 (0.0219)	0.8803 (0.0217)
p-value	BASE vs. LLE, $p = 0.0056$ ISO, $p = 0.0137$ LEM, $p = 0.0204$	PCA vs. LLE, $p = 0.0072$ ISO, $p = 0.0213$ LEM, $p = 0.0142$

Table 3.4: Mean (SD) of the volumes (in mm^3) in the left hippocampus in the baseline images of the atlas library of 110 images used to assess optimal methods and parameters.

	Control (n=19)	AD (n=36)
Manual (SD)	2749 (273)	2054 (424)
Automated (SD)	2722 (249)	2066 (387)
Man vs Auto mean of difference (p-value)	27 (p=0.19)	-12 (p=0.14)
SD of differences	129	150

Table 3.3 compares the mean Dice's similarity index (SD) obtained by selecting atlases with manifold learning and using the BASE and PCA methods. The results show that all 3 manifold learning selection methods significantly outperform $(p < 0.05)$ the BASE and PCA method.

Table 3.4 shows the mean (SD) of the manual and automated hippocampal volumes. The automated volumes were computed using Locally Linear Embedding with the optimized parameters. The mean (SD) of differences between the manual and automated hippocampal volumes by baseline diagnostic group was 27 (129) mm^3 (automated<manual) for controls and -12 (150) mm^3 (automated>manual) for AD subjects. In order to test the validity of my method, I compare the proposed method to a state-of-the-art method for hippocampus segmentation based on a similar atlas library approach (Leung et al., 2010). Using the same library of 110 hippocampus images and optimal parameters defined in Leung et al. (2010), a similar leave-one-out method is performed. The mean Dice's similarity index was 0.8955 (0.0172) compared to 0.9077 (0.0211) in my method. Even though these values differ by 0.01 point only, the difference is statistically significant (p<0.001). Figure 3.2 plots the volume correlation between the manual segmentation and my automatic segmentation method. The volume differences between manual segmentation and automatic segmentation are similar to zero-mean random noise. Figure 3.3 shows an example of segmentation obtained with my method.

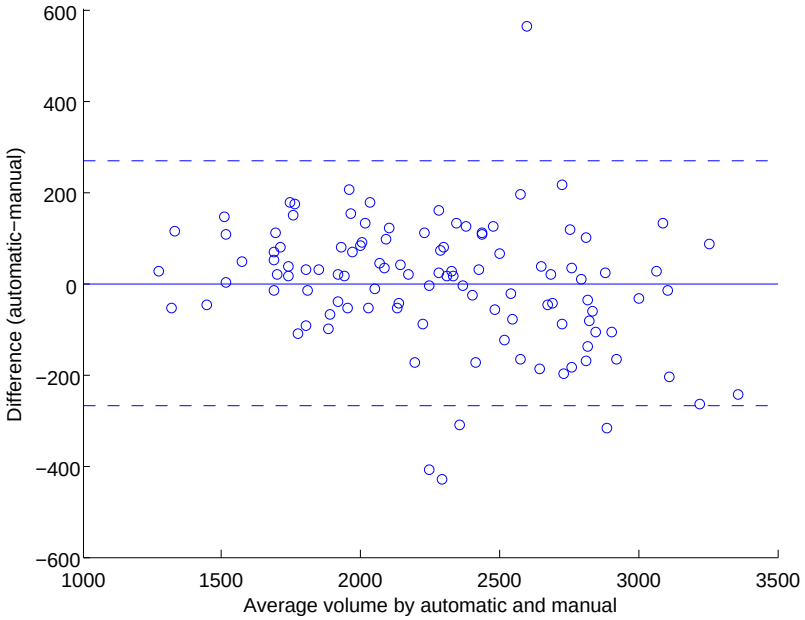

Figure 3.2: Bland-Altman plot. Each point corresponds to an hippocampal segmentation. The difference between automatic and manual estimates is plotted against their average. The solid horizontal line corresponds to the average difference, and the dashed lines are plotted at average +/-1.96 standard deviations of the difference.

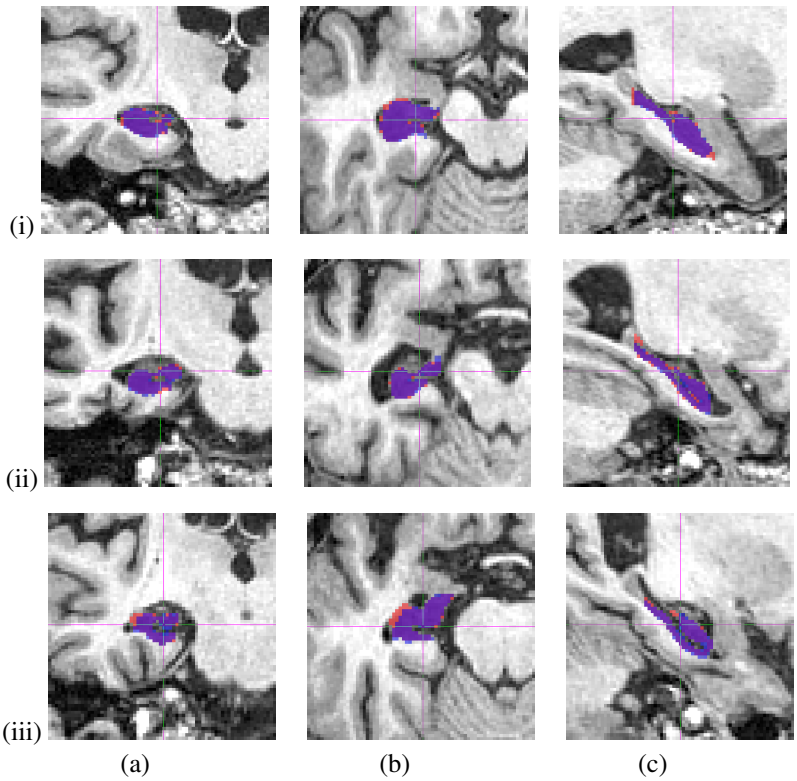

Figure 3.3: Hippocampal segmentation: automated (blue) vs manual (red). Overlapping area in purple. Row: (i) High case (Dice = 0.9398), (ii) Typical case (Dice = 0.9073), (iii) Low case (Dice = 0.8614). Column: (a) Coronal view, (b) Sagittal view, (c) Axial view.

Table 3.5: Mean (SD) of the volumes (in mm^3) in the left hippocampus in the baseline images of the labelled ADNI data set of 30 images for method validation.

	Control (n=10)	MCI (n=10)	AD (n=10)
Manual (SD)	2531 (336)	2331 (410)	1994 (478)
Automated (SD)	2642 (360)	2334 (431)	2018 (387)
Man. vs Auto. mean of diff. (p-value, paired t-test)	-111 (p=0.33)	-3 (p=0.47)	-24 (p=0.29)
SD of differences	168	155	130

Table 3.6: Effect size.

	ES_{AD}	ES_{MCI}
Manual (SD)	-1.124	-0.490
Automated (SD)	-1.614	-0.720

Overall, these results show that registering atlases that have been selected by manifold learning (i.e. selection in the lower-dimensional space) produces accurate and robust segmentation in the framework of multi-atlas based segmentation and gives better results compared to atlas selection without manifold learning (i.e. selection in the high-dimensional space). Also, given this data set of atlases, Locally Linear Embedding gives significantly better results than Isomap and Laplacian Eigenmaps.

3.5.2 Results from method validation

I use Locally Linear Embedding with the optimal parameters found in §3.5.1 to generate automatic segmentation of the 30 ADNI subjects. The mean (SD) Dice's similarity indexes of the left hippocampus segmentations of the baseline ADNI images are 0.887 (0.020) for controls, 0.886 (0.025) for MCI, 0.878 (0.038) for AD and 0.883 (0.028) across the three groups. These are summarized in Figure 3.4. The difference in accuracy compared to the previous experiment can be explained by the fact that the atlases and the 30 ADNI subjects belong to different data sets. Also the high shape variability and the possible presence of cysts in the hippocampus can explain lower scores in AD subjects. Table 3.5 shows the means (SD) of the manual and automated hippocampal volumes. The mean (SD) of differences in the manual and automated hippocampal volumes by baseline diagnostic group are -111 (168) mm^3 for controls, -3 (155) mm^3 for MCI, and -24 (130) mm^3 for AD subjects with automated volumes higher than manual volumes in all the three groups. Overall, the mean (SD) of differences in the manual and automated hippocampal volumes is -45 (154) mm^3. I also calculate the effect size $ES_{AD} = (\mu_{AD} - \mu_C)/\sigma_{AD}$ and $ES_{MCI} = (\mu_{MCI} - \mu_C)/\sigma_{MCI}$ in Table 3.6, where μ_C, μ_{MCI}, μ_{AD} are the average volumes in the control, MCI and AD groups respectively, and σ_{MCI}, σ_{AD} are the standard deviations in the MCI and AD groups respectively.

3.6 Conclusions

I compared Isomap, Locally Linear Embedding and Laplacian Eigenmaps for the selection of atlases to use in multi-atlas segmentation of the hippocampus of normal controls and patients with Alzheimer's disease in MR images. I found that Locally Linear Embedding generated the best hippocampal segmentation ($DS = 0.9077$) on a leave-one-out experiment using this data set of 110 atlases. The mean volumes and SDs of the generated segmentations were similar to those produced using manual segmen-

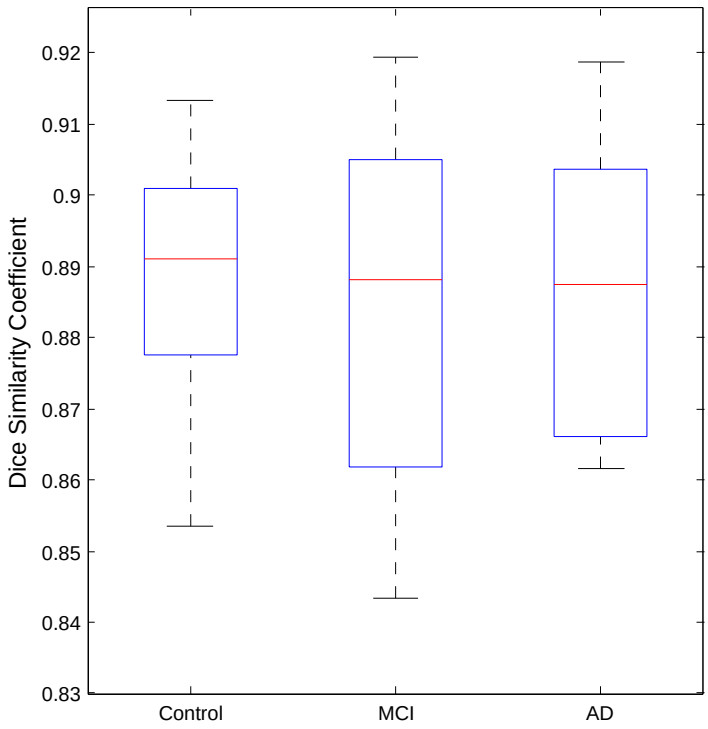

Figure 3.4: Average Dice's similarity index for NC, MCI and AD group obtained by fusing top 7 atlases with STAPLE. Atlases were selected with manifold learning.

tation. Overall, the mean difference between my automated volumes and the manual measurements was 7.5 mm^3 or around 0.01% of the mean of all volumes. I found good accuracy of my method on unseen data, achieving a mean Dice's similarity index of 0.883 (0.028) when comparing the automated and manual segmentations of a set of 30 subjects (10 AD, 10 MCI and 10 controls). Overall, the mean (SD) of differences in the manual and automated hippocampal volumes was 45 (154) mm^3 with manual<automated.

My results are consistent with those in Awate et al. (2012). They found that large number of k_d-nearest neighbours leads to higher Dice's similarity index for large database size M and that Dice's similarity index decreases as k_d approaches the value of M. In this study, the Dice's similarity index quickly rises to a maximum when the number of k_d-nearest neighbours increases for all the manifold learning techniques. The Dice's similarity index then gradually declines as the number of k_d-nearest neighbours increases.

Not only is the choice of manifold learning important but also the parameters used to compute the embedding. For instance, most studies have represented the embedding with 2 or 3 dimensions as it enables spacial visualization of the embedding. However the optimal embedding could have been of higher dimensions. Indeed, in this study, I found that the best results arose when using 11 dimensions. Also all manifold learning techniques presented in this chapter require the choice of a neighbourhood size either for the calculation of the geodesic distance in Isomap, or reconstructing a data point with its closest points in Locally Linear Embedding or Laplacian Eigenmaps. The choice of the optimal dimension and best parameters is often made empirically.

The results showed that selection of atlases with manifold learning is beneficial in the framework of multi-atlas based segmentation. The optimal accuracy can be found by fine tuning the manifold learning process. It also turned out that this atlas data set of hippocampi can be described by 3 main modes of variation regardless of the manifold learning technique used.

I found that Locally Linear Embedding gave best results for this data set of the hippocampus but it might not yield optimum results for a different anatomical structure. There is no consensus on which manifold learning technique to use for a given data set. A legitimate question that arises is which manifold learning algorithm is best suited for which data set. As demonstrated in this study, different manifold learning techniques produce different low-dimensional embeddings even for the same data set. This can be explained by the fact that the cost function to optimize associated with a manifold learning technique differs from one method to another.

The lower Dice's similarity index obtained when segmenting the 10 AD subjects from the ADNI data may also illustrate the issue of manifold sampling. Since the manifold is directly learned from points (i.e. images) in the data set, the sampling of the manifold is highly correlated with the density of points in the high-dimensional space. For example, if certain areas in the high-dimensional space are too sparse, the resulting manifold is likely to be a poor approximation of the true manifold structure. Since the atlas data set did not contain any MCI subjects, the manifold derived from this atlas data set is not representative of a population containing NC, MCI and AD subjects. It would have been preferable to derive a manifold from NC only subjects in the atlas data set to segment the 10 NC from the ADNI data set, and similarly for the 10 AD in the ADNI data set.

An important aspect in manifold learning is the metric used to relate pairs of images in the high-dimensional space. The most commonly used metrics are based on voxel intensity such as the Euclidean distance, cross correlation or mutual information. Similarly to Gerber et al. (2010) and Hamm et al. (2010), I used a metric derived from non-rigid transformation. In theory, the metric used should reflect the information relating pairs of images (Pless, 2004; Souvenir and Pless, 2005). However, there is currently no research investigating the influence of the metric on the resulting embedding. In the future, I am planning to compare the effects of several metrics such as the geometric median and the geodesic estimation proposed by Fletcher et al. (2009) and Avants and Gee (2004) respectively on low-dimensional embeddings.

I have obtained one of the best accuracies reported to date for automated hippocampal segmentation when compared with gold standard manual segmentations from a set of 30 randomly chosen subjects (10 AD, 10 MCI and 10 controls) from ADNI. My Dice's similarity index is equal to 0.88 with the previous highest Dice's similarity indexes (N=number of hippocampi in the study) being 0.86 (N=14) (Fischl et al., 2002), 0.83 (N=60) (Heckemann et al., 2006a), 0.81 (N=100) (Pohl et al., 2007), 0.86 (N=54) (Barnes et al., 2008b), 0.87 (N=30) (Chupin et al., 2008), 0.88 (N=5) (Gousias et al., 2008) (from a cohort of 2 year old children), 0.86 (N=40) (Morra et al., 2008), 0.85 (N=30) (Powell et al., 2008), 0.86 (N=40) (van der Lijn et al., 2008), 0.83 (N=550) (Aljabar et al., 2009), 0.89 (N=160) (Collins and Pruessner, 2010), 0.89 (N=30) (Leung et al., 2010), 0.89 (N=120) (Lötjönen et al., 2010) and 0.85

(N=364) (Wolz et al., 2010a). Our intra-rater variability corresponds to a Dice's similarity index of 0.96. Comparing this to the results from using my automatic method with different training and test data (0.88) suggests that the method has not been over-trained, and that there is potential to improve it further.

Overall, my technique is most similar to that reported by Wolz et al. (2010a). However it fundamentally differs in the following ways: (i) Wolz et al. (2010a) used a similarity measure derived from voxel intensities, whereas I used a metric derived from registration. (ii) I embedded target images using the out-of sample extension instead of embedding all images in a single manifold. This method effectively scales with the number of atlases and not the number of images to segment. (iii) I used STAPLE as a fusion method, whereas statistical voxel classification and graph cuts was used in Wolz et al. (2010a).

I developed a suitable method for segmenting large data sets by extending the manifold with an out-of-sample image. Indeed, in my method: (i) the low-dimensional manifold learned from the space spanned by the set of atlases, (ii) the average atlas M and (iii) the registrations between the atlases and M are precomputed and stored, thus making my method very computationally efficient. I only need to perform one non-rigid registration between M and a new unseen target image x to select its most similar images from the atlases. This method is therefore scalable and extremely computationally efficient, making it suitable for segmenting large data sets and for clinical use.

3.7 Summary

Several manifold learning techniques have been applied to various datasets. But the key question of which manifold learning and what parameters to choose for a given dataset remains unanswered. In many cases, it is not possible to predict how a manifold learning algorithm will perform for a given dataset. When using manifold learning for atlas selection in the framework of atlas-based segmentation, inaccuracy in segmentation can come from multiple sources including image registration, the label fusion technique used or the atlas selection by the manifold learning technique. It is therefore important to limit the impact of manifold learning in contributing to segmentation errors. In this chapter, 3 manifold learning techniques were trained and tuned to segment a dataset of hippocampus. The same image registration and label fusion algorithm was used. LLE was found to perform best on the dataset of hippocampus. As a result, this manifold learning is subsequently chosen to segment structures at risk in CT images in Chapter 4.

Chapter 4

Validation of clinical acceptability of atlas-based segmentation for the delineation of organs at risk in head and neck cancer

4.1 Introduction

Intensity-modulated radiotherapy (IMRT) enables normal tissue sparing by allowing better conformal dose distribution in head and neck cancer tissue. This technology requires the accurate delineation of several target volumes (TVs) and surrounding organs at risk (OARs). This delineation is typically performed manually by trained experts on computed tomography (CT) or magnetic resonance (MR) images and sometimes complemented with functional imaging techniques such as positron emission tomography (PET) (Kruser et al., 2009; Newbold et al., 2006). This process may need to be repeated multiple times during radiotherapy treatment to accommodate to tumor response and physiological changes in the patient.

In practice, manual contouring is time-consuming and labor intensive, especially for large TVs and irregular OARs. It is also subject to large inter-rater variability (Hong et al., 2004; Jeanneret-Sozzi et al., 2006), despite universally accepted delineation guidelines (Grégoire et al., 2006, 2003; Sjöberg et al., 2013). Mean volume variations of up to 50% were reported in parotid delineation across three radiation oncologists on CT images (Geets et al., 2005). Further investigations showed that the effects of inter-rater variability in delineating OARs has a significant dosimetric impact (Nelms et al., 2012). In addition, the range of inter-rater variability has been found to be greater in some cases than errors due to positioning and organ motion (Weiss and Hess, 2003). Consequently, the development of accurate and reproducible automatic segmentation method is crucial to allow clinicians to focus on other aspects of patients treatment.

Recently, automatic atlas-based segmentation methods have shown promising results in segmenting head and neck CT images (Stapleford et al., 2010; Young et al., 2011). Different methods have been developed based on either a single-patient atlas (Commowick and Malandain, 2007), a population-based average atlas (Commowick et al., 2008), or multiple atlases (Teguh et al., 2011). Multi-atlas methods have been shown to yield better results than single atlas methods (Sjöberg et al., 2013; Teguh et al.,

2011). For the fusion of multiple atlases, the "Simultaneous Truth and Performance Level Estimation" (STAPLE) algorithm (Warfield et al., 2004) has been used in several studies to generate contours in the head and neck region (Han et al., 2008; Stapleford et al., 2010; Teguh et al., 2011). Since the introduction of the original STAPLE algorithm, other segmentation methods that build upon it have been proposed to take into account the similarity between the atlases and the image to segment. In particular, Cardoso et al. (2013) developed the "Similarity and Truth Estimation for Propagated Segmentations" (STEPS) algorithm. In STEPS, atlases are locally ranked based on their similarity with the image to segment using the locally normalized cross-correlation. For a local region to segment, only the top ranked atlases for that region are used during the fusion process. In contrast, all atlases carry the same global weight in STAPLE. STEPS has previously been validated on brain structure segmentation (Irani et al., 2013; Ma et al., 2014), and has been shown to perform better than STAPLE. This is is line with the fact that local fusion strategies outperform global methods (Artaechevarria et al., 2009).

A standard evaluation of accuracy has been the direct comparison of manual and automatic segmentations using overlap measures such as the Dice similarity coefficient (DSC) (Dice, 1945). However, the accuracy of automatic methods as measured this way is limited by the degree of inter-rater variability in manual contouring. In the presence of such variability, even an algorithm that performs as well as an expert can not be expected to achieve total agreement with manual segmentations. Furthermore, it is possible that an automatic segmentation does not resemble the gold standard, but is still acceptable for use in radiotherapy planning. This judgment can not reliably be made based on overlap measures, and an expert rater decision is required.

Automated methods can reduce physician contouring time by up to 30-40% as seen in studies of head and neck cancer (Sjöberg et al., 2013), and also reduce the inherent inter-rater variability in volume delineation (Stapleford et al., 2010). The improvement in time and consistency are valuable only if segmentation accuracy is not undermined. Assessing the accuracy of automatic segmentation is a challenging task and manual editing is usually required to achieve clinically acceptable results (Stapleford et al., 2010; Young et al., 2011). Nevertheless, the workload of manual editing can be significantly shorter than manual contouring (Sjöberg et al., 2013).

In this study, I compare STAPLE against STEPS in producing accurate segmentations for radiotherapy planning. Both algorithms are used to segment the following OARs in head and neck cancer: the brainstem, the spinal canal, the left and right parotids, the optic chiasm, and the eyes. The accuracy of both algorithms was measured using the Dice similarity coefficient (DSC) (Dice, 1945). In addition to accuracy, I measure the clinical acceptability of each automatic method. To account for the variability in overlap measures, manual contours and automatic segmentations produced by STAPLE and STEPS were graded on a 3-point scale for clinical acceptability in a blind experiment by 3 distinct trained physicians. The comparison through blindly obtained grades of manual and automatic segmentations represents a novel approach for their evaluation. Traditional evaluation has been to directly compare manual and automatic segmentations using the DSC. Although a high DSC should guarantee clinical acceptability, a lower DSC does not necessarily mean that an automatic segmentation is not clinically useful. To our

knowledge, methods classifying segmentations for clinical acceptability on a point scale by expert raters have not been published before. Time gain by using automatic segmentation was also assessed.

4.2 Related publications

- **Hoang Duc A.K.**, Eminowicz G., Mendes R., Wong S.L., McClelland J., Modat M., Cardoso M.J., Mendelson A.F., Veiga C., Kadir T. and Ourselin S.: Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. Medical Physics. *In press*.

4.3 Materials and methods

4.3.1 Overview

First, 6 OARs were delineated by two radiation oncologists in a dataset of 100 patients with head and neck cancer on computed tomography (CT) images. Each patient in the dataset was automatically segmented with both the STAPLE and STEPS algorithms using those manual contours. DSC was then used to measure the accuracy of the automatic segmentations. Second, 3 separate and distinct trained physicians graded the manual and automatic segmentations generated by both methods into one of the following 3 grades in a blind experiment: clinically acceptable without modification, fulfilling universal delineation guidelines (Cefaro et al., 2013) for radiotherapy planning (grade A), reasonably acceptable for clinical practice upon manual editing (grade B) and not acceptable (grade C). DSC for the STEPS algorithm and for each grade was then calculated. Last, STEPS segmentations graded B were selected and given to one of the 3 physicians who manually edited them to grade A. Editing time were recorded.

4.3.2 Atlas dataset

The atlas dataset consisted of N = 100 planning CT images of patients with different diagnoses of head and neck cancer. These were cases treated with IMRT at the radiotherapy department for any head and neck cancer diagnosis (squamous cell cancer and adenocarcinoma), including post-operative and primary radiotherapy with diagnoses including pharyngeal, laryngeal, oral cavity, unknown primary and maxillary sinus cancer. Staging ranged from T2N0M0 to T4N3M0.

Each CT image was acquired using a General Electric RT CT scanner and was composed of 100 to 205 slices (2.5mm thick) containing 512 x 512 pixels each. All patients were scanned head-first supine with their head blocked by an anatomical cushion and an individual thermoplastic mask. Our study involved 100 patients: a first radiation oncologist contoured 43 patients, and a second distinct radiation oncologist contoured the remaining 57 patients. For each patient, six OARs in the head and neck region were manually contoured for radiotherapy purposes. This included the brainstem, the spinal cord, the parotids (left/right), the optic chiasm, and the eyes. The eyes volume comprises the left and right side of the orbits, lenses and optic nerves. This grouping was deliberate. Since those structures are small, spreading only a couple of axial slices, and are generally delineated successively one side after the other, it was coherent to group them under a single label. Also, this was done to align the time scoring of the eyes with the time scoring of the other OARs (i.e. brainstem, the spinal cord, the parotids (left/right),

the optic chiasm). Some traditional OARs (i.e. lymph nodes, mandible) used in head and neck planning were not investigated. Indeed, not all traditional OAR segmentations were available for all patients. In a large amount of cases, the lymph nodes (either left or right), the mandible or the vocal cord were not available to us for this study. As a result, we only considered the OARs that were available for every patient which were the brainstem, the spinal canal, the left and right parotids, the optic chiasm, and the eyes.

4.3.3 Atlas-based segmentation

A registration algorithm is used to create automatic segmentations of regions of interest for a new image by transforming existing segmentations of the corresponding structures in existing images. Those automatic segmentations are then combined into a single consensus using a fusion algorithm.

4.3.3.1 Registration algorithm

A leave-one-out experiment was used in which each patient (referred to as a target) in the dataset was automatically segmented using the remaining atlases. A registration algorithm (Modat et al., 2010) was used to deform the atlases onto the target image space. The target image space is defined as the space of the patient to segment. The manual contours were then mapped onto the target using the resulting transformation from registration and fused with either the STAPLE or STEPS algorithm to yield estimated segmentations. The registration first determined an affine registration using translation, rotation and scaling. The affine registration used a symmetric approach of the block-matching algorithm developed by (Ourselin et al., 2001). A multi-level non-rigid registration step using free-form deformations with a cubic B-spline control point parameterization (Rueckert et al., 1999) was subsequently applied. The locally normalized cross-correlation was used as a similarity measure. The control point spacing was 5 voxels in all directions and a bending energy penalty term was used to regularize the deformation. The time to perform affine and non-rigid atlas registration onto a patient target image is about 45 min using a regular CPU.

4.3.3.2 Fusion using the STAPLE and STEPS algorithms

The STAPLE and STEPS algorithms are both based on an expectation-maximization (EM) framework. The framework starts with computing an estimate of the ground truth using a simple segmentation method. Based on this initial guess, it is possible to calculate the performance of each individual label. In the expectation step (E-step), labels are combined to estimate the true segmentation depending on their performance. In the maximization step (M-step), given an estimate of the true segmentation, the performance values of each labels are re-assessed and is maximized. In general, the performance is dependent on certain parameters and the M-step is used to find the parameters which maximize the performance of each label, while in the E-step, the estimate of the true segmentation is improved based on these parameters. In STAPLE, each segmentation is weighted globally depending upon their estimated performance level in the E-step and the sensitivity and specificity of each label is calculated in the M-step. In STEPS, the sensitivity and specificity is only calculated in areas where each classifier is considered an expert by the LNCC ranking strategy. This results in a 2 step performance estimation that

decouples the two sources of error: one based on the LNCC image similarity metric observation characterizing the non uniform registration accuracy and shape differences, and the other step characterizing the specificity and sensitivity of each classifier when compared with the consensus classification. Due to the local nature and smoothness of the metric, the similarity between the images is described on a smooth voxel by voxel basis, enabling a voxel by voxel ranking with reduced discontinuity effect. The raw HU units were used to compute the LNCC metric.

When a dataset of atlases is available, it is best to select the most similar atlases to the target when using STAPLE rather than using the whole dataset (Aljabar et al., 2009; Leung et al., 2010). To apply STAPLE in this study, I followed the method in Hoang Duc et al. (2013) based on manifold learning for atlas selection as the method showed consistently good results in selecting atlases. In Hoang Duc et al. (2013), three dimensionality reduction techniques (Isomap, Locally Linear Embedding and Laplacian Eigenmaps) were compared for the selection of atlases to use in multi-atlas segmentation. This study also investigated the optimal number of atlases to fuse for each technique. Optimal results were obtained by choosing the best 7 atlases using locally linear embedding. Therefor, for each target, the best 7 atlases where selected using the locally linear embedding method (Roweis and Saul, 2000). In contrast, STEPS does not require an explicit atlas selection as the algorithm already integrates a local ranking scheme. In this study, the whole dataset was registered to the target. Once all registrations are done, the top 7 ranked registered atlases for each local region (i.e. a patch of 5×5 voxels) to segment were used in the fusion process. As a result, STEPS does not require an atlas selection strategy but more registrations need to be performed than in STAPLE. Indeed, STEPS requires as many registrations as the size of the atlas dataset. The time to perform atlas fusion is about 5 min using a regular CPU. So total time to obtain an automatic segmentation (registration and fusion) is about 50 min.

4.3.4 Evaluation

The first objective was to compare the STAPLE against the STEPS algorithm in producing accurate segmentations. DSC between manual contouring and the two automatic segmentation methods was reported. It is defined as $D(U,V) = 2|U \cap V|/(|U| + |V|)$, where $|U|$ (resp. $|V|$) is the number of voxels in the automated (resp. manual) region. Its value ranges from 0 to 1, where 0 means no overlap, and 1 signifies a perfect match.

4.3.5 Segmentation grading

The second objective was to assess whether the STAPLE and STEPS algorithms could produce segmentations as clinically relevant as manual contouring. All segmentations were imported into a treatment planning system (Varian Eclipse version 11) and graded by a trained physician. Three distinct physicians, with the same level of expertise as the two radiation oncologists, graded in a blind experiment manual and automatic segmentations using one of the following 3 grades:

- Grade A: the segmentation is clinically acceptable and satisfies universal OAR delineation guidelines (Cefaro et al., 2013) and can be used as created for radiotherapy planning.
- Grade B: the segmentation is reasonably acceptable but needs some manual editing. Some contour

lines need to be corrected to meet universal guidelines.

- Grade C: the segmentation does not meet universal guidelines. Some slices show gross mis-delineation that cannot be attributed to segmentation variability.

On this scale, grade A is considered higher than grade B, and grade B higher than grade C. The 3 distinct physicians graded manual and automatic segmentations in a random order. To reduce bias from assessing the same structure multiple times, associated automatic and manual segmentations were graded at least 1 week apart. The first physician graded the 6 OARs of 100 patients. Due to time constraint, the second and third physicians could only grade the 6 OARs of 50 and 30 patients respectively. Comparison between grades of manual and automatic segmentations by the 3 trained physicians is used as an indicator of clinical acceptability. Although radiation oncologists contours were graded by 3 distinct trained physicians, this does not imply that one expert rater was better than another. A total of 1200 automatic and 600 manual segmentations were graded (1200 = 6 OARs x 100 patients $\times$ 2 and 600 = 6 OARs $\times$ 100 patients).

4.3.6 Manual editing time

The third objective was to quantify manual contouring time saved by using the STEPS algorithm. When patients were originally contoured for radiotherapy treatment, contouring time was not recorded. In order to estimate this contouring time and to keep manual contouring to an acceptable level, one of the 3 trained physicians re-contoured the OARs of 5 patients and the time was recorded. Those 5 patients were chosen to be representative of the whole dataset by an external researcher. Time reported for the eyes volume was the aggregated time to contour the component parts. For each OAR, the physician was given 15 randomly selected STEPS segmentations graded B and edited them to grade A. Editing time was recorded. A brush to push in/out the contour lines, freehand and eraser tools were used for contouring and editing.

4.4 Results

4.4.1 STAPLE vs STEPS

The DSC are reported in Figure 4.1. Significant improvements can be seen when using the STEPS algorithm on large structures such as the brainstem, spinal canal and left/right parotid compare to the STAPLE algorithm. Using a Wilcoxon rank-sum test, STEPS segmentations yielded significantly higher DSC than STAPLE segmentations (all $p < 0.001$) for those structures. For smaller structures such as optic chiasm and the eyes, the difference are not significantly different ($p > 0.300$ and $p > 0.170$). The DSC for those structures are significantly lower compare to larger ones. This can be explained by their size, where even small voxel mis-classification in the automatic segmentation will result in large DSC discrepancy. Figure 4.2 shows some examples of manual, STEPS and STAPLE segmentations of the brainstem, the spinal canal, and the parotids (left/right). The clinical acceptability of our method could not have been reliably determined with the DSC, and verification by means of separate trained physicians was required.

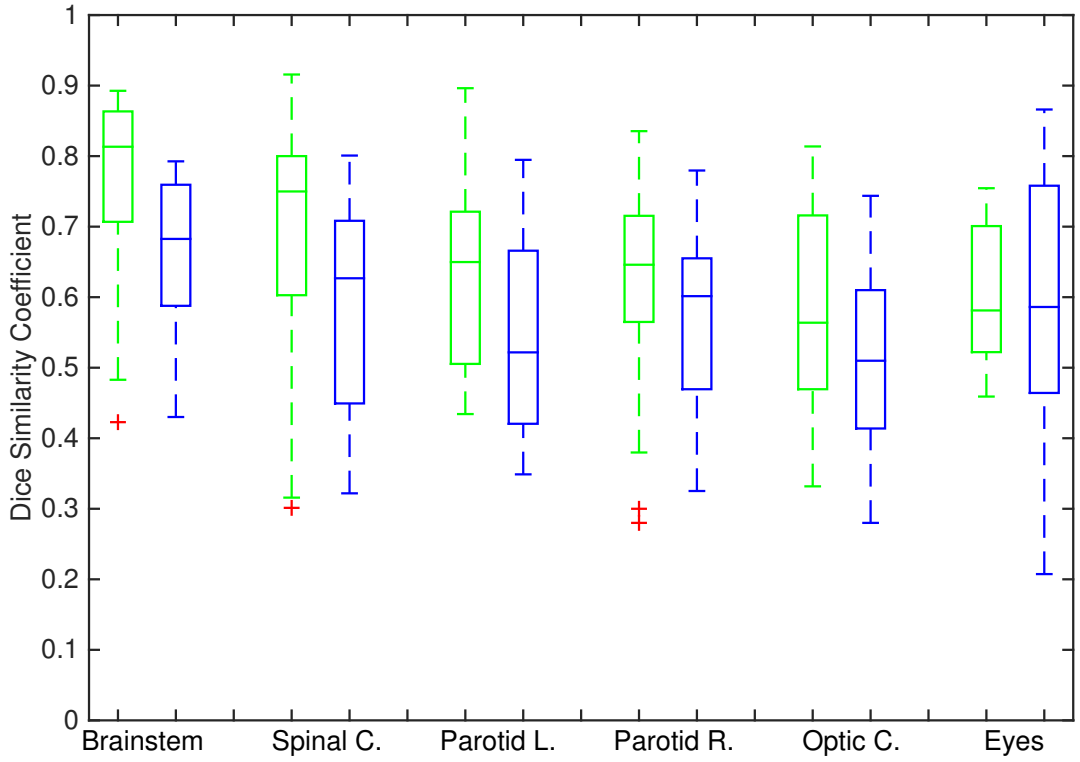

Figure 4.1: Dice similarity coefficient of the STEPS (green) and STAPLE (blue) algorithm against manual contouring.

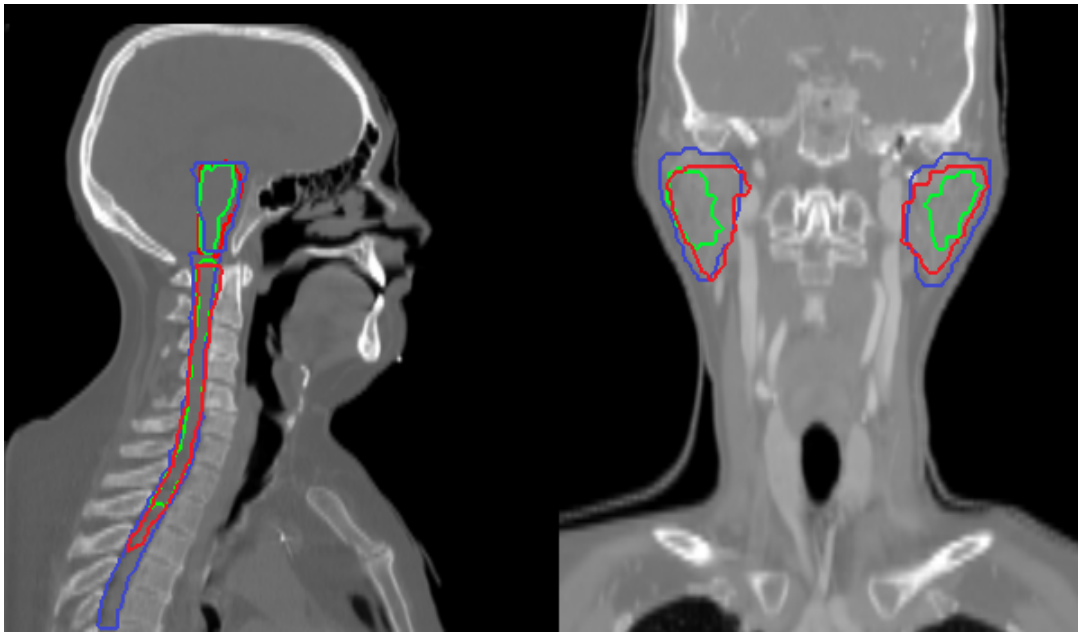

Figure 4.2: Examples of manual (blue), STEPS (red), and STAPLE (green) segmentations of the brainstem, spinal canal and parotids (left/right).

4.4.2 Grading

Results of grading by the 3 trained physicians are shown in Figure 4.3. A surprising number of manual contours for the eyes and optic chiasm were graded B and C, corresponding to high inter-rater variability. This is consistent across the 3 trained physicians. This may be due to the poor contrast of those areas in CT images. Manual and STEPS segmentations of the parotids (left/right) and the optic chiasm were given similar grades by 2 trained physicians. The third physician, except for the left parotid, drew similar conclusion. When similar grades where given, a Wilcoxon signed-rank test did not show any significant difference for those OARs (all $p > 0.100$). For the brainstem and the spinal canal, STEPS segmentations were overall graded similarly as well. In some cases, STEPS segmentations of those OARs were graded higher than manual segmentation and those differences were statistically significant ($p < 0.010$). In contrast, STEPS segmentations the eyes were graded significantly lower ($p < 0.005$).

Overall, STAPLE segmentations were graded significantly lower than both manual and STEPS segmentations (all $p < 0.01$), except for the optic chiasm and the eyes ($p > 0.273$ and $p > 0.382$).

Figure 4.4 shows the grade distribution of STEPS, which gave the best results out of the two automatic methods, and manual segmentations. Only distribution from the trained physician who graded all 100 patients is shown. It can be noted that a substantial number of STEPS segmentations of the spinal canal (27 cases) and the eyes (30 cases) were graded lower than their associated manual contours, and I offer some explanation. The well defined boundaries of the spinal canal make it one of the easier OARs to segment for an expert rater, but atlas-based methods were seen to suffer from two key problems there. High neck flexion confounded registration in 10 cases, and discrepancies in the length of the lower part segmented in the atlas set (vertebrae below C1) caused failure in 17 more. No atlas-based method can overcome such discrepancies, and they must be fixed by standards in the templates used. For the eyes, since the structures involved are small, a slight deviation in the automatic segmentation will inevitably result in some manual editing being required.

Across all OARs, STEPS was observed to outperform STAPLE and produce segmentations graded as well as or better than manual contours with a rate of 83%. A one sided confidence interval based on the t-statistic places the true rate above 80% with 95% confidence.

4.4.3 Dice similarity coefficient and clinical acceptability

To examine the relationship between acquired grades and DSC, I calculated the DSC between clinically acceptable (grade A) manual contours only and the STEPS segmentations graded A, B and C. Only the segmentations from the physician who graded all 100 patients are examined. Results are presented in Figure 4.5. Using a Wilcoxon rank-sum test, STEPS segmentations graded A did not yield significantly higher DSC than STEPS segmentations graded B. The median DSC was also seen to vary significantly between OARs, for instance, the median DSC of the left/right parotids were significantly different from all other regions (all $p < 0.020$). Therefore, it may not be meaningful to compare segmentation quality between different regions using this measure. For all OARs, DSC of STEPS segmentations graded C were significantly lower (all $p < 0.005$) compared to segmentations graded A and B. Since STEPS segmentations graded A and B yielded similar DSC, the clinical acceptability of my method could not

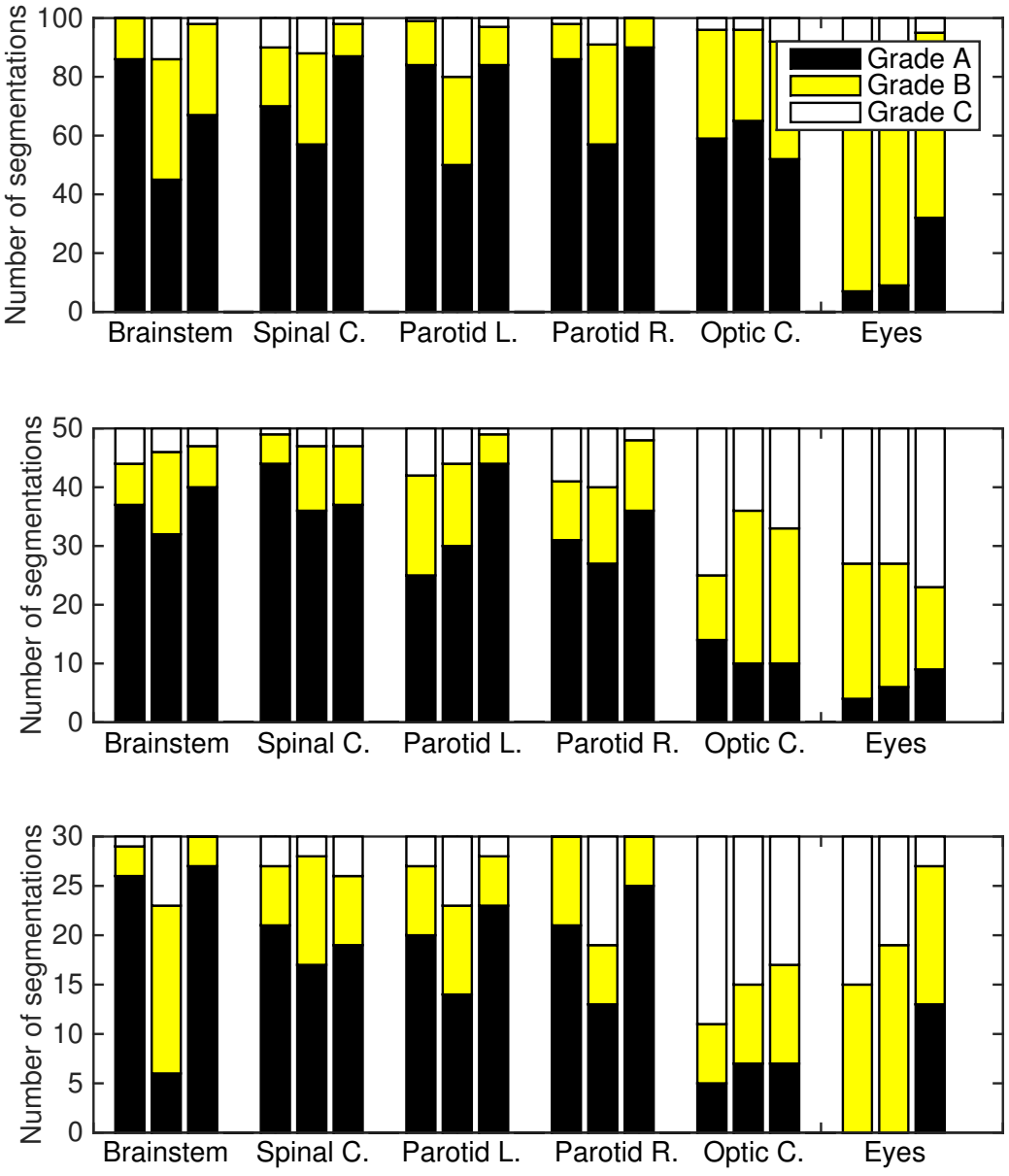

Figure 4.3: Grading of manual and automatic segmentations by 3 distinct trained physicians. Each graph represents grading done by a physician. For each OAR: STEPS = left bar, STAPLE = middle bar, Manual = right bar. Grade A: clinically acceptable, no editing required. Grade B: reasonably acceptable, some editing required. Grade C: not acceptable.

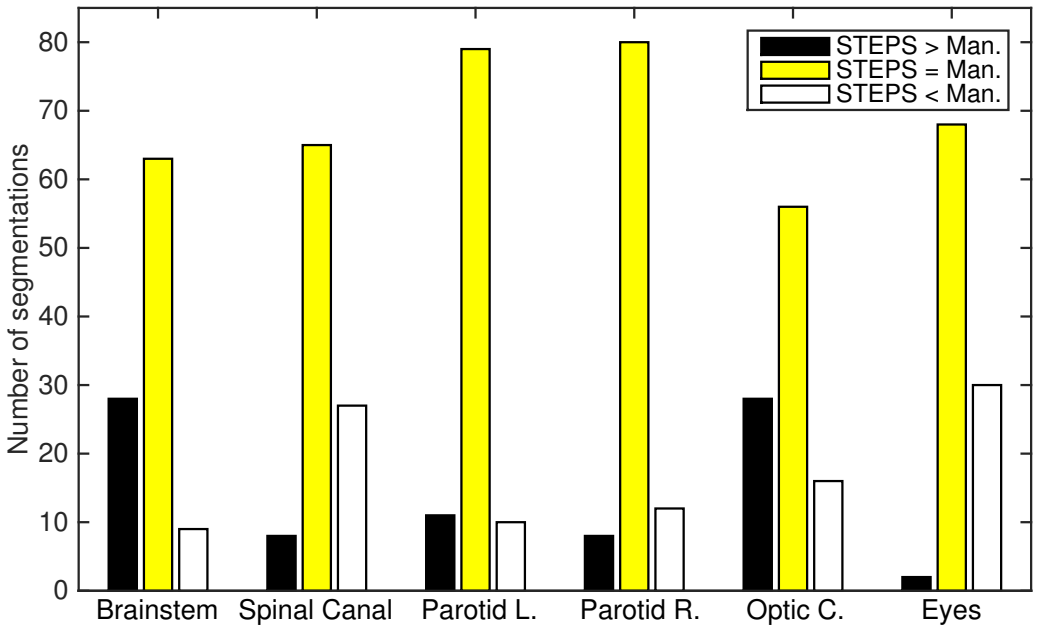

Figure 4.4: Grade distribution of automatic and associated manual segmentations. STEPS $>$ Man.: STEPS segmentation has a higher grade than its associated manual contour. STEPS = Man.: STEPS and manual segmentations have the same grade. STEPS $<$ Man.: STEPS segmentation has a lower grade than its associated manual contour.

	Auto. without editing vs. Man scratch	Auto. with editing vs. Man scratch
Brainstem	95.16%, $p < 10^{-3}$	69.46%, $p < 10^{-3}$
Spinal Canal	91.77%, $p < 10^{-3}$	64.50%, $p < 0.005$
Parotid Left	92.53%, $p < 10^{-3}$	67.42%, $p < 10^{-3}$
Parotid Right	94.58%, $p < 10^{-3}$	70.10%, $p < 10^{-3}$
Optic Chiasm	92.43%, $p < 10^{-3}$	66.80%, $p < 10^{-3}$
Eyes	95.28%, $p < 0.01$	28.26%, $p < 0.005$

Table 4.1: Relative gain (%) in segmentation time. PTvalues are the results of the Wilcoxon rankTsum test.

have been reliably determined with DSC, and verification by means of a separate trained physician was required.

4.4.4 Time scoring

Figure 4.6 shows the time taken to obtain a grade A result using the STEPS algorithm with manual editing, without it, and using fully manual contouring. Using the Wilcoxon rank-sum test, these results demonstrate that STEPS yielded significant time saving, even when automatic segmentation needed editing. Time saved is relatively lower for the eyes, these being a grouping of 6 different structures, the trained physician spent a significant amount of time switching between editing tools, which added to the effective editing time. Time gained and p-values are reported in Table 1. Time gained is calculated using the following ratio: (grading time + editing time) / (manual contouring time) if the automatic segmentation needed editing, and (grading time) / (manual contouring time) if the automatic segmentation didnt need editing.

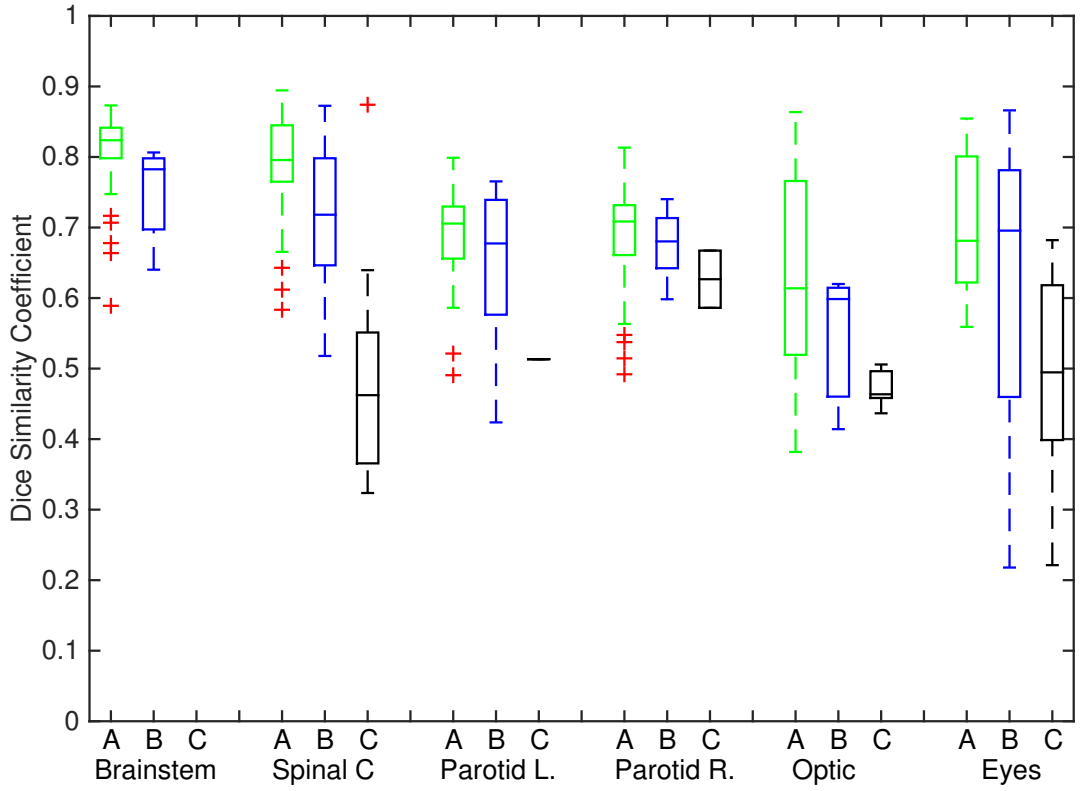

Figure 4.5: Dice similarity coefficient of STEPS segmentations graded A (green), graded B (blue) and grade C (black) versus manual contours graded A. Only the segmentations from the physician who graded all 100 patients are shown.

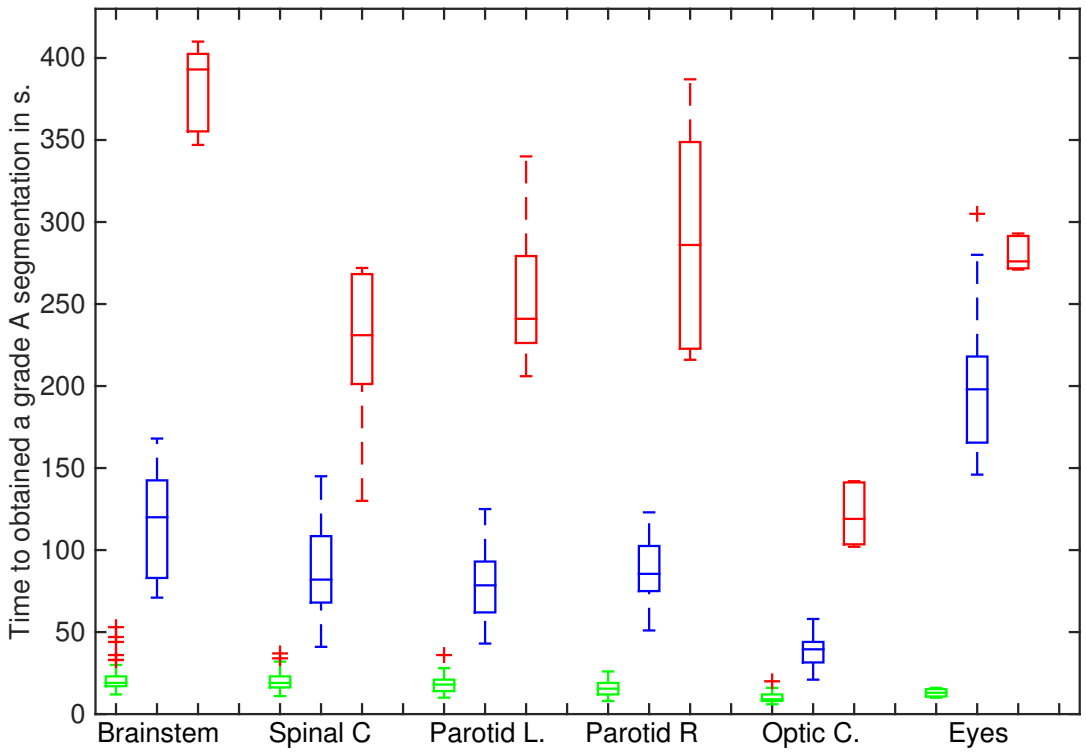

Figure 4.6: Time in seconds to obtain a grade A segmentation using STEPS algorithm without (green) or with (blue) manual editing and with fully manual contouring (red).

4.5 Discussion

In this study, the STAPLE and STEPS algorithms used multiple manual contours to generate the most likely segmentation using information from the radiation oncologists. Inter-rater variability is one of the most challenging issues in IMRT and is a motivation for the development of methods that improve consistency. The results showed the advantages of STEPS over STAPLE in segmenting OARs in head and neck cancer. In summary, DSC from STEPS were higher compared to DSC from STAPLE for the brainstem, spinal canal and left/right parotids. This showed that the local combination strategy introduced in STEPS outperform the global fusion method in STAPLE. In addition, STEPS produced segmentations that were as clinically acceptable as manual contouring for structures such as the brainstem, spinal canal, parotids (left/right), and optic chiasm. In contrast, STEPS segmentation grades of the eyes were lower than grades from manual contouring. DSC reported in this study compare well with DSC reported in the literature (0.78 and 079 for the brainstem and parotids gland in (Teguh et al., 2011), 0.75 and 0.72 in (Daisne and Blumhofer, 2013)). Across all OARs, I found a reduction in time of 61% and 93% on average when STEPS segmentation did and did not respectively require manual editing. This time gain was superior to numbers previously reported in the literature (40% in (Daisne and Blumhofer, 2013), 26% in (Stapleford et al., 2010), and 47% in (Chao et al., 2007)).

The better results generated by STEPS over STAPLE are in line with findings in the literature. In Cardoso et al. (2013) , the robustness and accuracy of STEPS were evaluated on a database of cross-sectional and longitudinal brain MRI scans. In that study, STEPS performed better that STAPLE. STEPS has also been successfully used in other papers Irani et al. (2013); Ma et al. (2014) to segment MR images. However only our studies and the one from Cardoso et al. (2013) directly compared the performance of STEPS and STAPLE and further investigation will need to be done across various range of image modalities to check if this statement holds.

A standard evaluation approach in radiotherapy has been to directly compare manual and automatic segmentations using the DSC. However this study demonstrated that the DSC does not reliably reflect clinical acceptability of an automatic segmentation. Although a high DSC should mean clinical acceptability, a lower DSC does not necessarily mean that an automatic segmentation is not clinically useful. It may then be counterproductive to use a particular minimum DSC as a threshold for clinical acceptance of an automatic method, even if this is calibrated for a particular OAR.

Atlas-based segmentation is highly dependent on the similarity between the underlying atlas and the patient (Rohlfing et al., 2005). In this study, the failure in delineating the spinal canal in some cases could be due to multiple factors: a) bad performance of the registration algorithm around that area, b) lack of images in the atlas dataset with the same overall spinal morphology, c) labeling discrepancies in the manual segmentation of the spinal canal (i.e. discrepancies in the length of the lower part segmented in the atlas set (vertebrae below C1)), and d) patient head and neck position in the scanner when images are acquired. Different segmentation strategies based on either a single patient atlas, a population-based average atlas, or multiple atlases have intrinsic limitations due to large deformations of normal anatomy that cannot be corrected with registration algorithms. Importantly, when thinking about applying auto-

mated segmentation, clinical concern arises due to abnormal anatomy in patients developing head and neck cancer. My dataset included a variety of cases including some with bulky tumors, and results with our method were still comparable to manual contouring for the brainstem, spinal canal, left/right parotid and optic chiasm across the cohort. In any case, automatic segmentations should always be checked and corrected if necessary by an expert before planning.

Starting contouring from an existing template (either automatic or manual) may have influenced the trained physicians perception of gold standard. In general, relatively minor editing to the segmentations was performed and the lack of modifications may be attributed to the fact that the segmentations closely resembled physicians definition of gold standard. However, this scenario represents the common clinical situation of verifying contours from less experienced clinicians, where relatively minor modifications are usually made overall.

Finally, there are some limitations to this study. Limitations include the small number of OARs edited and manually contoured to measure time cost and the lack of assessment of intra-rater variability. However, these limitations should not affect the conclusion drawn as the significant p-values are all below 0.01 despite a wide confidence interval. In addition, this study did not include TVs. Multi-modality imaging is often used to improve the visibility of TVs by co-registering CT with MR or PET images. Unfortunately, I did not have access to any imaging modalities other than CT. I note that atlas-based methods perform well when the shape of the target is well represented in the dataset of atlases, which is rarely the case in radiotherapy as tumors have no predefined shape.

4.6 Conclusions

The STEPS algorithm shows better performance than the STAPLE algorithm in segmenting OARs for radiotherapy of the head and neck. It is clinically useful and can considerably save time for clinicians in contouring OARs for radiotherapy planning. Even though automatically generated segmentations should always be checked and approved by an expert before radiotherapy planning, the STEPS segmentation method was found to be comparable to manual contouring for the brainstem, spinal canal and left/right parotid.

Chapter 5

Generating synthetic CT images from MR scans for radiotherapy treatment of the head and neck

5.1 Introduction

The calculation of the dose distribution is a critical step in radiotherapy planning as the irradiation of tumour tissue needs to be maximized while the irradiation to organs at risks (OARs) needs to be minimized during treatment. In particular, the head and neck region contains a large number of OARs, and therefore requires a high level of accuracy in dose calculation. This calculation requires the knowledge of the linear attenuation coefficient of irradiated tissues (Hoppe et al., 2010). Tissues with a large attenuation coefficient quickly attenuate radiation beam, whereas tissues with a small attenuation coefficient are relatively transparent to the beam. Computed tomography (CT) imaging has been the modality of choice for providing such information in form of CT numbers expressed in Hounsfield units (HU) (Parker et al., 1979). HU are defined by the following equation: $HU_{tissue} = [(\mu_{tissue} - \mu_{water})/\mu_{water}] \times 1000$, where μ is the linear attenuation coefficient of the medium. For example, $HU_{air} = -1000$, $HU_{water} = 0$, $HU_{tissue} \in [100, 300]$ and $HU_{bone} > 700$. HU values can be converted into relative electron density by using a look-up table within the treatment planning system. The look-up table can be described by two linear fits. Figure 5.1 represents a look-up table of CT number against physical density. The calculation of dose distribution can subsequently be performed with the knowledge of this density (Seco and Evans, 2006).

Recently, radiotherapy planning solely based on magnetic resonance (MR) imaging without the use of CT imaging has gained popularity. MR imaging provides superior soft tissue contrast which is beneficial for segmentation in the head and neck region (Evans, 2008). It has been shown that manual delineations of the parotids on MR scans have higher inter-observer agreement than those on CT scans (Rasch et al., 1997). Margins added to target volume delineation to account for uncertainties could be reduced by using MR imaging resulting in less radiation to normal tissues and a reduction in treatment toxicity (Rasch et al., 1999; Roach III et al., 1996). For instance, in the case of prostate cancer, volume delineation on MR and CT images were compared by Roach III et al. (1996). The bony anatomy on

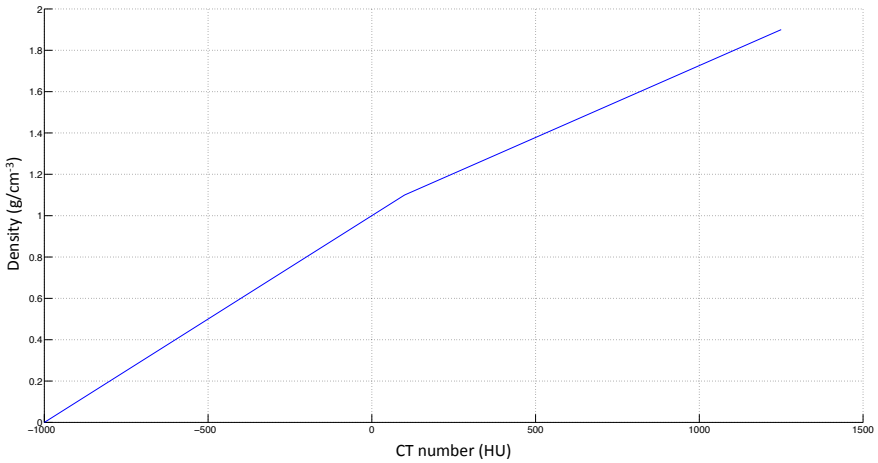

Figure 5.1: A look-up table to convert from CT number to relative electron density used by a treatment planning system for dose calculation. The look-up table is generally derived from the CT scan of a phantom containing a number of medium of know density.

the images was matched and the measured prostate volumes compared. The mean prostate volume was found to be 32% larger on average on CT images. However, MR has yet to replace CT imaging in clinical practice. Indeed, CT imaging provides higher anatomical accuracy compared to MR imaging, the latter being subject to geometric distortions caused by magnetic inhomogeneities, non-linear gradients, susceptibility and chemical shifts (Doran et al., 2005; Stanescu et al., 2010). In addition, MR imaging does not directly provide electron-density information for dose calculation.

There is no direct mapping between MR intensity and the underlying electron density of a tissue. As a result, several approaches have been proposed to generate CT-like images and derive electron density from MR images (Dowling et al., 2012; Karotki et al., 2011; Lee et al., 2003; Uh et al., 2013). Those CT-like images are commonly referred to as pseudo CT or synthetic CT in the literature. One proposed method has been to segment MR images into distinctive structures (typically bony structure, soft tissue, and air) and assign corresponding bulk CT intensities to each structure. Figure 5.2 from Lee et al. (2003) presents an example of a synthetic CT image generated using bulk assignment. Bulk electron density assignment has shown dosimetric results similar to those based on electron density provided by real CT images (Karotki et al., 2011; Lee et al., 2003). State-of-the-art methods use a spatial mapping between MR and CT images, similar to an atlas-based propagation (Dowling et al., 2012; Uh et al., 2013), to estimate probable CT intensities at each voxel location of a given MR image. In Dowling et al. (2012), a single atlas composed of a paired CT/MR image was used to generate a synthetic CT image from a given target MR image in prostate radiation therapy. The MR atlas was registered to the target MR image and the resulting deformation was applied to the CT atlas yielding a synthetic CT image. Overall, the dose distributions calculated on the synthetic CT image were in close agreement with the original doses. Another approach is to use multiple pairs of CT/MR atlases to create a synthetic CT image by fusing the deformed CT atlases after registration to the target. In Uh et al. (2013), averaging of voxel intensities with different number of atlases was used to combine CT atlases of the brain. This study suggested that synthetic CT images created from multiple deformed atlases are more suitable for treatment planning

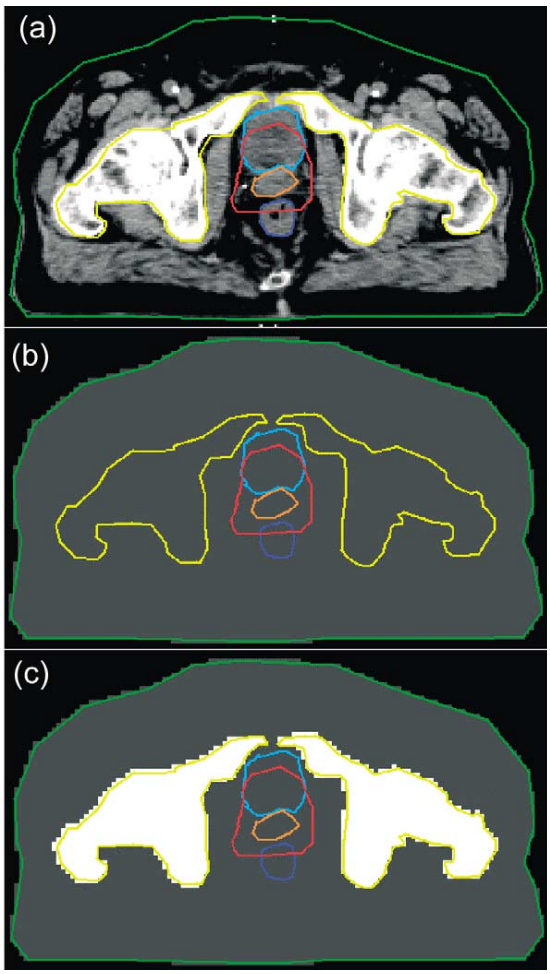

Figure 5.2: Corresponding slices of (a) CT, (b) water and (c) bone and water bulk assigned images. The bulk-assigned values were equivalent to water and average bone value. Orange, red, dark blue and light blue outlines are GTV, PTV, rectum and bladder, respectively. Yellow and green outlines are of bone and patient contour. Figure from Lee et al. (2003).

than those from a single atlas or bulk electron density assignment. Nevertheless, atlas-based methods are prone to registration uncertainty, and heavily deformed anatomy in patients developing head and neck cancer increases the risk of registration errors.

In this chapter, the method presented by Burgos et al. (2013) that alleviates the registration uncertainty is used to synthesize an electron density map from a given target MR image. It is based on atlas propagation with the additional consideration of morphological similarity between patients. This morphological similarity, based on an image similarity measure, is applied so that the more morphologically the atlases are, the higher the weight they carry in the fusion process (Cardoso et al., 2012). The method employs aligned CT/MR pairs of images from multiple patients to propagate CT atlas intensities onto the target MR image in a voxel wise manner. It showed good result in creating an attenuation correction map for PET/MR scanners (Burgos et al., 2013). Since head and neck data show tremendous variability across patients, the use of morphological similarity could be highly beneficial during the fusion process. To the best of my knowledge, using atlas-based propagation for generating a synthetic CT image of

the head and neck area using morphological similarity represents a novel approach in electron density mapping for dose calculation.

5.2 Methods

5.2.1 Overview

A synthetic CT image providing electron density information was generated from a target MR image by deforming pairs of CT/MR atlases onto that target image. First, for each patient in the dataset, the CT image was aligned to the MR image to create a CT/MR atlas. Second, the MR atlases were registered to the target MR image. The resulting deformations were then applied to the CT atlases to create multiple deformed CT images. Third, those CT images, aligned with the target MR, are combined using a fusion strategy based on a spatially varying weighted averaging and atlas ranking to create a synthetic CT image. For evaluation, the image similarity between the real and synthetic CT images was compared using the mean absolute error. In addition, difference in dose distribution was calculated by replacing the real CT image in the treatment planning system with the synthetic CT image.

5.2.2 Data

The dataset consisted of $N = 30$ pairs of planning CT and T2 weighted planning MR images of patients with different diagnoses of oropharyngeal cancer. All patients were diagnosed with stage T2 to T3 tumors. They were sampled from a clinical trial aiming to evaluate the utility of functional magnetic resonance imaging (diffusion, dynamic contrast enhancement, spectroscopy and blood oxygenation level dependent contrast) for the detection of residual nodal disease following chemo-radiotherapy to head and neck squamous cell cancer. Images from both modalities were acquired during the same day. Patients were placed in a supine position on a rigid couch with their head blocked by an anatomical cushion during image acquisition. Each CT image was composed of 125 to 170 slices, and each slice contained 512×512 voxels. The CT images were acquired on a GE Wide Bore 16 slice system with contrast injection and an imaging resolution of $0.977 \times 0.977 \times 2.50$ mm. MR images were acquired using a T2 axial sequence with a Siemens Avanto scanner. They were composed of 60 or 61 slices, with 4 cases containing less than 50 slices. Each MR slice contained 256×256 voxels. The image resolution was $0.703 \times 0.703 \times 3.30$ mm. MR images were corrected for intensity non-uniformity following a non-parametric non-uniform intensity normalisation procedure (Sled et al., 1998).

5.2.3 CT/MR atlas creation

The method used in this chapter requires the creation of a pair of CT/MR atlas for each patient in the dataset. CT images in my dataset were acquired for treatment planning. MR images were acquired for pre-treatment planning. As a result, the field of view (FOV) in CT images included the entire head and neck region extending up to the superior part of the lungs. In contrast, FOV in MR images only encompassed the target volume which included the top of the brainstem up to vertebrae C1 or C7. Due to the acquisition from different imaging modalities, the CT and MR images needed to be aligned to create a paired CT/MR atlas. An affine followed by a non-rigid cubic B-spline registration (Modat et al.,

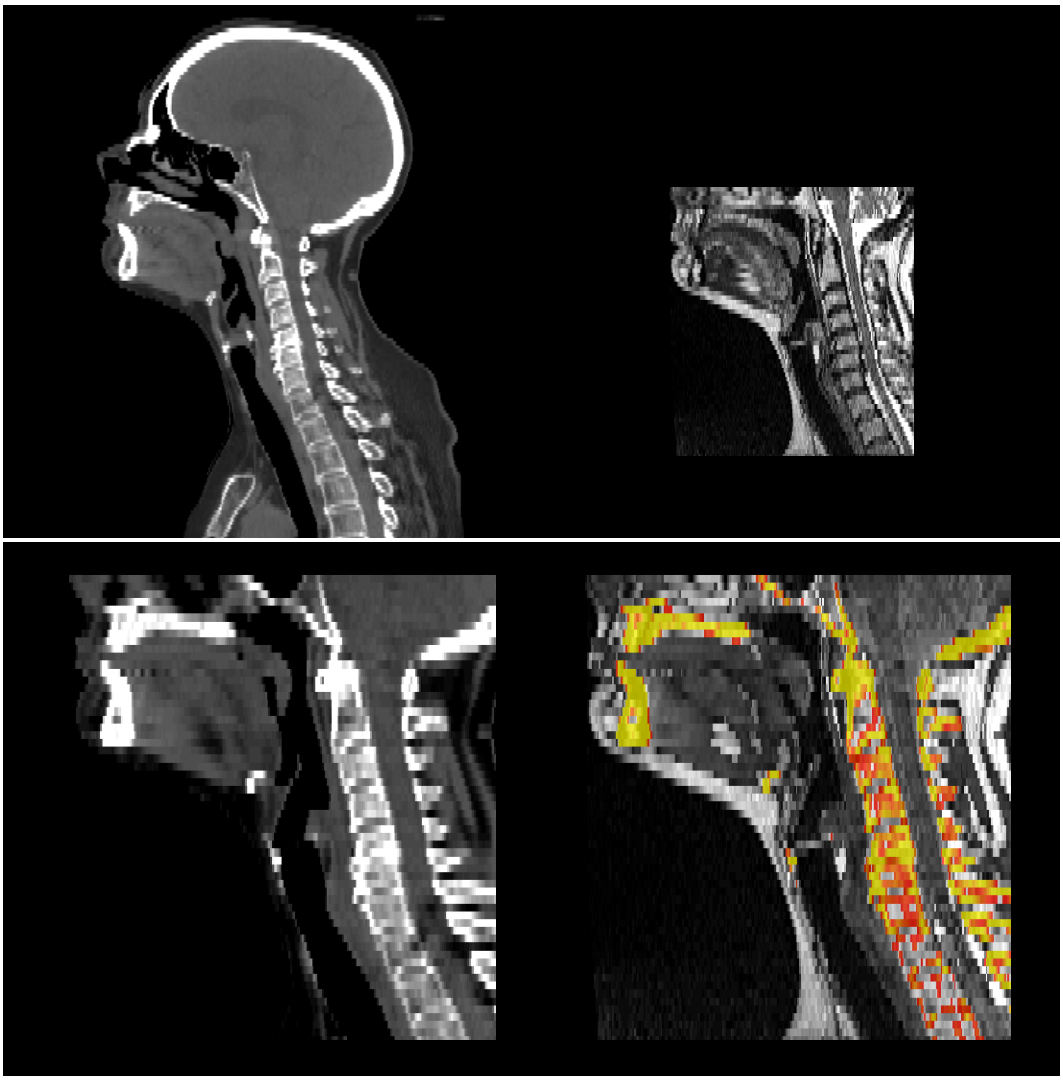

Figure 5.3: Top left: planning CT image. Top right: planning MR image. Bottom left: CT image in the space of the MR image. Bottom right: CT image overlaid on the MR image.

2010) was applied to align the CT to the MR image. The normalized mutual information (NMI) was used as a similarity measure along with a control point spacing of 10 mm. All the results from the registration were assessed visually, and no critical mis-alignement were reported. Figure 5.3 presents an example of a CT and corresponding MR image. In an ideal situation, the planning CT and MR image would have been acquired with the same FOV so that the two modalities contained the same or comparable geometric information. However, using retrospective data, this is unlikely to be possible as there is yet no clinical reason to acquire a planning MR image with an extended FOV due to time and cost. In addition, acquisition of larger FOV requires longer acquisition time to obtain the same image resolution, and is prone to additional artefacts such as patient motion.

5.2.4 Construction of a synthetic CT image

A leave-one-out experiment was performed in which a synthetic CT image was constructed for each target MR image in the dataset using the remaining atlases. First, all MR atlases are registered to the target MR image using a symmetric global registration followed by a cubic B-spline parametrised non-

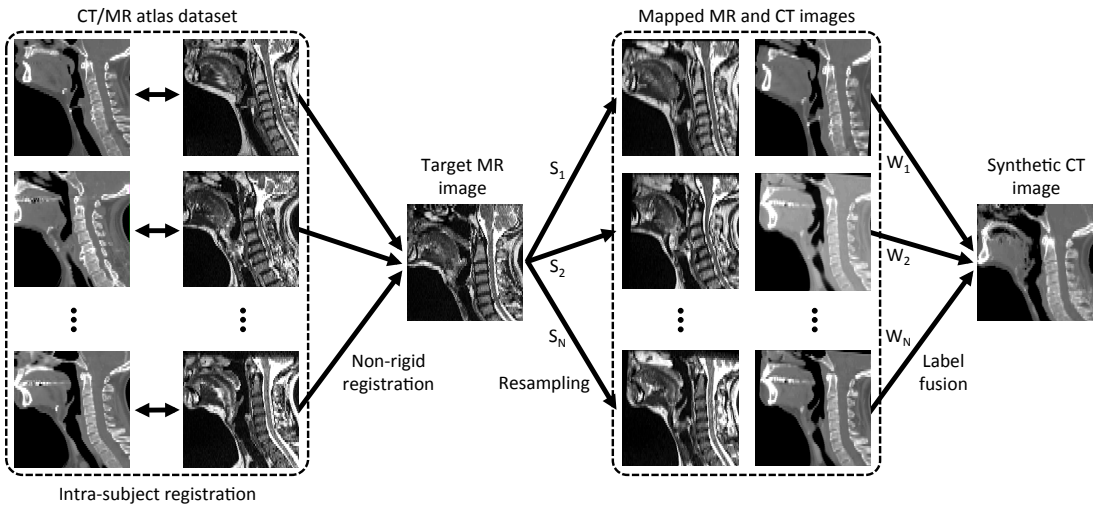

Figure 5.4: Illustration of CT synthesis for a given MRI image. All the MR images in the atlas dataset are registered to the target MR image. The CT images in the atlas dataset are then mapped using the same transformation to the target MR image. A local image similarity measure (S) between the mapped and target MRIs is converted into weights (W) to generate the synthetic CT image.

rigid registration, using NMI as a measure of similarity (Modat et al., 2010). A control point spacing of 5 mm was employed to account for morphological differences in inter-subject registration. Second, CT atlases were mapped onto the target using the resulted transformation that aligned the subject's corresponding MR atlas to the target MR image. Through this process, we obtained a series of paired CT/MR images aligned to the target MR image.

A synthetic CT was then constructed by combining multiple CT atlases aligned to the target into a single consensus following the method presented in Burgos et al. (2013). In this method, the intensity of each synthetic CT voxel is obtained by a locally varying weighted averaging. For a given local region on the target MR image, local weights are calculated using local similarity between the target MR image and the MR atlases mapped onto it. Providing that the local image intensity similarity is a good approximation of the local morphological similarity between patients, it can be assumed that if two MR images are similar at a certain spatial location, the associated CT images will also be similar at this location (Cardoso et al., 2012). Figure 5.4 illustrates the process of generating a synthetic CT image. Details of the label fusion are provided next.

The local similarity measure used was the the convolution-based fast local normalised correlation coefficient (LNCC) proposed by Cachier et al. (2003). In the following, the target MR image is denoted by T^{MRI} and its corresponding unknown CT by T^{CT}. For each of the N atlases in the database, the registered MR and CT images of atlas n are denoted by A_n^{MRI} and A_n^{CT} respectively. The LNCC between T^{MRI} and A_n^{MRI} at voxel $\vec{v}$ is then given by:

$$LNCC_{n,\vec{v}} = \frac{\langle T^{MRI}, A_n^{MRI} \rangle_{\vec{v}}}{\sigma(T^{MRI})_{\vec{v}} \sigma(A_n^{MRI})_{\vec{v}}} \tag{5.1}$$

The mean and standard deviation at voxel $\vec{v}$ were calculated using a Gaussian kernel G_{σ_G} with

standard deviation σ_G through convolution:

$$\overline{T}_{\vec{v}} = [G_{\sigma_G} * T]_{\vec{v}} \qquad \sigma(T)_{\vec{v}} = \sqrt{\overline{T^2}_{\vec{v}} - \overline{T}^2_{\vec{v}}} \qquad \langle T, A \rangle_{\vec{v}} = \overline{T.A}_{\vec{v}} - \overline{T}_{\vec{v}} \cdot \overline{A}_{\vec{v}} \tag{5.2}$$

High LNCC values indicate a better local match between the two MR images. In addition, registered MR atlases were also ranked using a ranking scheme similar to the one proposed by Yushkevich et al. (2010) in order to compensate for registration inaccuracies, giving a larger weight to MR atlases better registered to the target MR image. The LNCC at each voxel were ranked across all images, with the rank being denoted by $K_{n\vec{v}}$. The ranks $K_{n,\vec{v}}$ were then converted to weights by applying an exponential decay function:

$$W_{n,\vec{v}} = e^{-\beta K_{n,\vec{v}}} \tag{5.3}$$

with $W_{n,\vec{v}}$ being the weight associated with the n^{th} subject image at voxel $\vec{v}$, and β is a coefficient influencing the repartition of the weights. Similarly to the label fusion framework suggested by Cardoso et al. (2012), an estimate of the target CT image can be obtained by a spatially varying weighted averaging. The weights $W_{n,\vec{v}}$ were used to reconstruct the target CT image T^{CT} at voxel $\vec{v}$ as follows:

$$T^{CT}_{\vec{v}} = \frac{\sum_{n=1}^{N} W_{n,\vec{v}} . A^{CT}_{n,\vec{v}}}{\sum_{n=1}^{N} W_{n,\vec{v}}} \tag{5.4}$$

Based on previous optimization experiments done by Burgos et al. (2013) on MR and CT brain images, σ_G and β were set to 3 and 0.5 respectively.

5.3 Evaluation

5.3.1 Synthetic CT accuracy

For each target MR image in the dataset, a synthetic CT image, called S^{CT}, was built using the proposed method. In addition, the single best MR atlas for each target was selected and used to generate a corresponding CT image, called best-atlas CT image B^{CT}. The best atlas was selected based on a global similarity measure: the normalised cross-correlation (NCC). This measure was computed over the entire image between each MR atlas mapped onto the target MR image to select the most similar atlas to the target. The NCC is defined as:

$$NCC_n = \frac{1}{V} \frac{\langle T^{MRI} - \overline{T^{MRI}}, A^{MRI}_n - \overline{A^{MRI}_n} \rangle}{\sigma(T^{MRI}) \sigma(A^{MRI}_n)} \tag{5.5}$$

where $\overline{T}$ is the mean and $\sigma(T)$ the standard deviation of image T.

The intensities of the S^{CT} and the B^{CT} images were compared to the real CT intensities R^{CT}. The metric employed to measure the synthesis error was the mean absolute error, defined as:

$$MAE(S^{CT}) = \frac{\sum_{\vec{v}} |S^{CT}_{\vec{v}} - R^{CT}_{\vec{v}}|}{V} \tag{5.6}$$

where V is the number of voxels in the region of interest. This cost function was estimated between the real CT image and the synthetic/best atlas CT image for every subject in the dataset. Moreover, compar-

isons of histograms between the real image and the synthetic/best atlas CT images were performed in order to detect possible bias.

5.3.2 Dose calculation

A dosimetric evaluation was performed to assess the suitability of using the synthetic CT image for dose calculation. The FOV of the synthetic CT image was the same as the FOV of the MR image. Since the synthetic CT image is missing some anatomy of the patient, the original CT scan could not have been replaced by the synthetic CT image in the treatment planning system for dose calculation. As a result, the original planning CT image was cropped to the same FOV of the MR image and was used as a ground truth for dose calculation. Doses were calculated for an IMRT plan using Varian Eclipse External Beam Planning System analytical anisotropic algorithm with a resolution of 2.5 mm. The individual IMRT plan for which each patient was treated with, including a 9-beam arrangement, monitor units and fluence maps, was applied. For each patient, dose calculation was estimated based on that individual original plan. In order to compare the proposed method with the bulk assignement method, a bulk-assigned CT image $Bulk^{CT}$ with a single value corresponding to water ($HU_{water} = 0$) assigned inside the patient body contour and the value of air ($HU_{air} = -1000$) outside was built. Dose calculation was then computed on the real (R^{CT}), synthetic (S^{CT}) and bulk-assigned ($Bulk^{CT}$) for comparison. Dose distribution for each case is referred to as D_R, D_S and D_{Bulk} respectively.

Dose calculation was done on 4 cases chosen to be representative of the dataset. The dose distributions were compared considering dose differences with a constraint of 2% prescribed dose. In addition, the dose volume histograms (DVHs) for 4 OARs (brainstem, spinal canal, left parotid, and right parotid) were computed. DVH is an histogram relating radiation dose to tissue volume. It is routinely used in clinical practice to assess if the plan is appropriate for the patient, by displaying information of dose delivered both to volume of interest. DVHs were computed using manual contours. The structures were delineated by the radiation oncologists as part of clinical practice.

5.4 Results

5.4.1 Comparison between synthetic CT and real CT images

Figure 5.5 shows an example of a synthetic CT image generated with the proposed method and a best atlas CT image. Overall, the synthetic CT images showed good visual similarities with the real CT images, especially in vicinity of bony structures. In contrast, the best atlas CT images displayed some missing anatomy as illustrated in Figure 5.5 due to the difference in field of view between the target and the best atlas. In addition, the synthetic CT images did not show dental artefacts that could be observed on the best atlas CT images. This can be explained by the local label fusion which can detect voxels of dental artefacts as outliers due to their high CT numbers and does not take them into account during fusion. The difference between the real and synthetic CT images was more pronounced around bony structures and surfaces between air and tissue. This is possibly due to bone/tissue and air/tissue interfaces not being precisely registered by the non-rigid transformation.

The MAEs were computed between the synthetic CT/best atlas CT images and the real CT images.

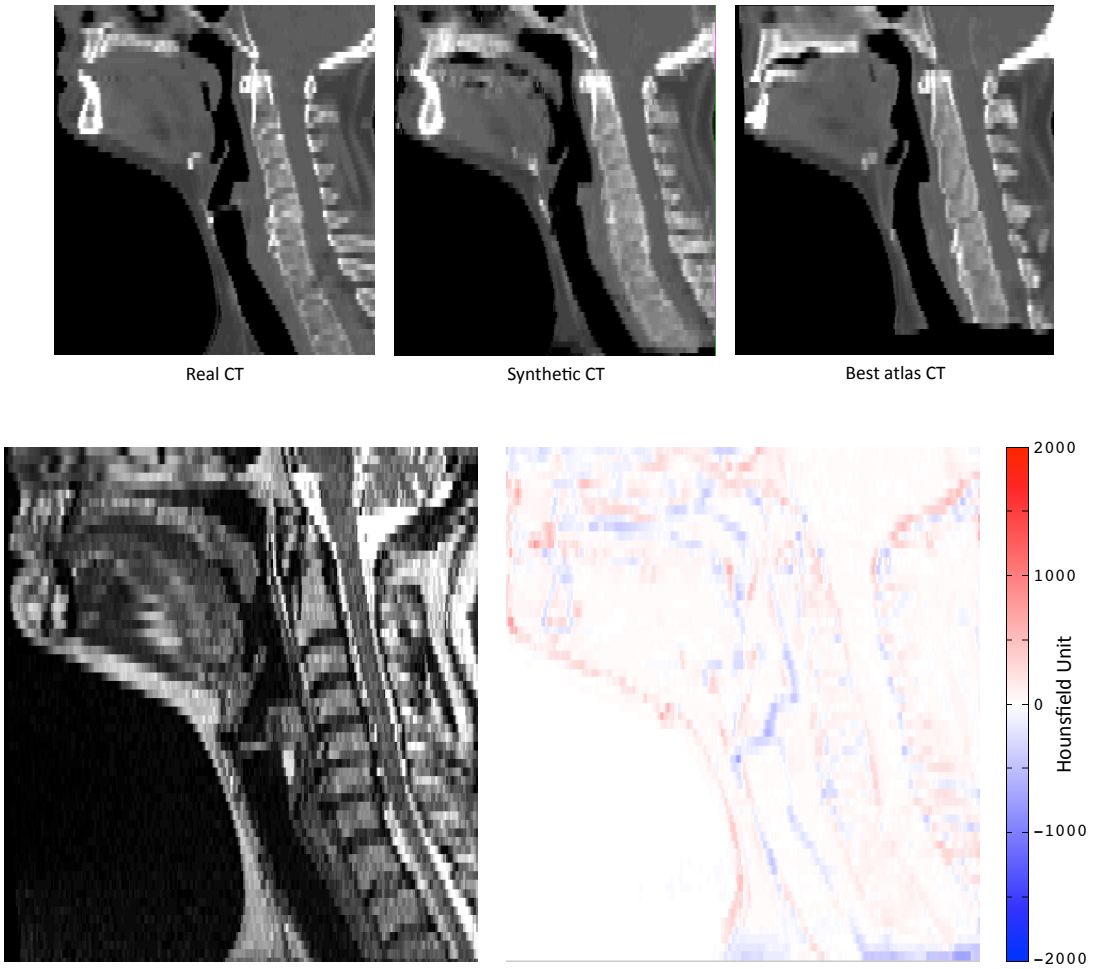

Figure 5.5: Top left: real CT image (R^{CT}). Top middle: synthetic CT image (S^{CT}). Top right: best atlas CT image (B^{CT}). S^{CT} shows good visual similarity with R^{CT}, especially in vicinity of bony structures. B^{CT} can have missing information. Bottom left: real MR image. Bottom right: difference between the real and synthetic CT image.

The mean (STD) MAE for the synthetic CT method was 131.8 ($\pm$ 31.5) HU, and 144.9 ($\pm$ 22.3) HU for the best atlas method. A paired t-test showed that this difference was significant ($p < 10^{-5}$). Figure 5.6 shows the distribution of the MAEs.

The average distribution of CT numbers in real, synthetic and best atlas CT images is shown in Figure 5.7. The synthetic CT images tend to underestimate CT numbers in the range -500 to 0 which correspond to air/tissue surfaces. This can be explained by the fact that the body contour in the real CT image is better defined compared to the body contour in the synthetic image.

5.4.2 Evaluation of dose calculated on synthetic CT images

The 4 cases on which dose calculation was done are presented in Figure 5.8. The dose difference between the synthetic/bulk-assigned CT image and the real CT image for a single case is presented in Figure 5.9. Overall D_S matches D_R well, whereas D_{Bulk} presents a global overestimation, meaning that more dose is delivered to tissue when using bulk-assigned CT images compared to real CT images. Compared to D_R, D_S was mostly different at the location of air/tissue surfaces. This is again due to the accuracy of the body contour in the synthetic image compared to the real CT image. The dose similarities were analysed

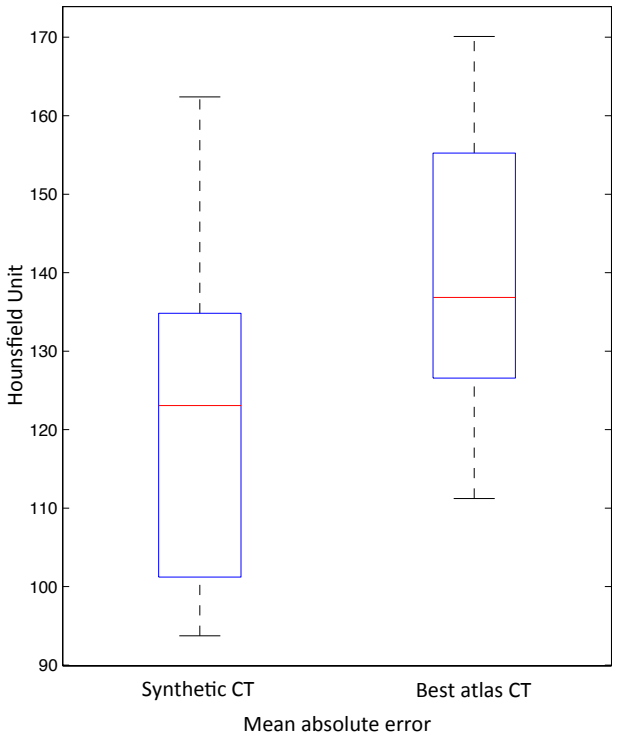

Figure 5.6: Boxplot showing the mean absolute error distribution. The central mark is the median and the edges of the box are the 25th and 75th percentiles. The whiskers extend to the most extreme data points.

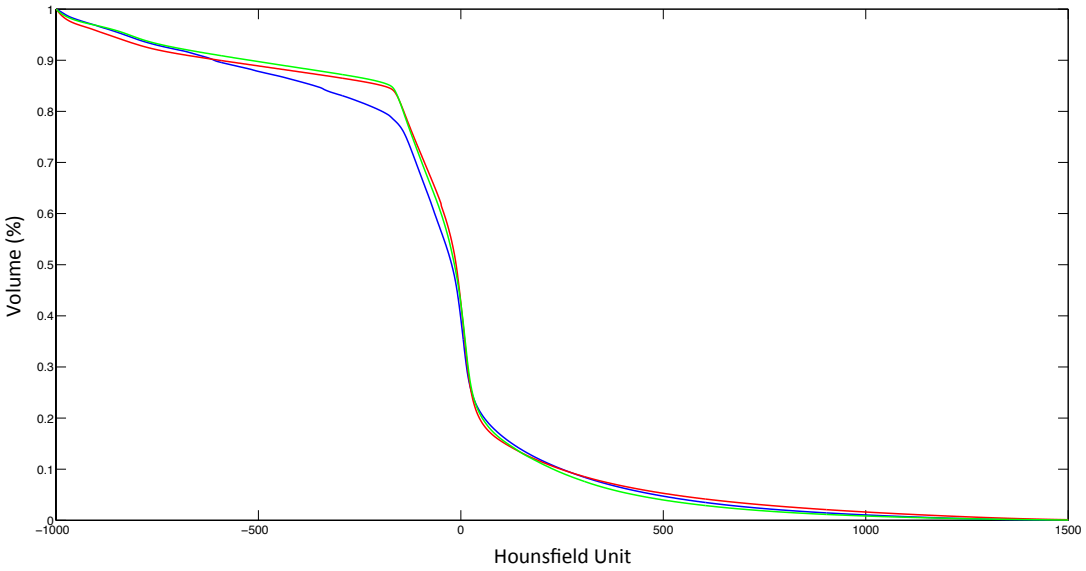

Figure 5.7: Distribution of CT number in real (red), synthetic (blue), and best atlas CT (green) images. Synthetic CT images tend to underestimate CT numbers in the range -500 to 0 which correspond to air/tissue surfaces. This can be explained by the fact that the body contour in real CT image is better defined compared to the body contour in synthetic image.

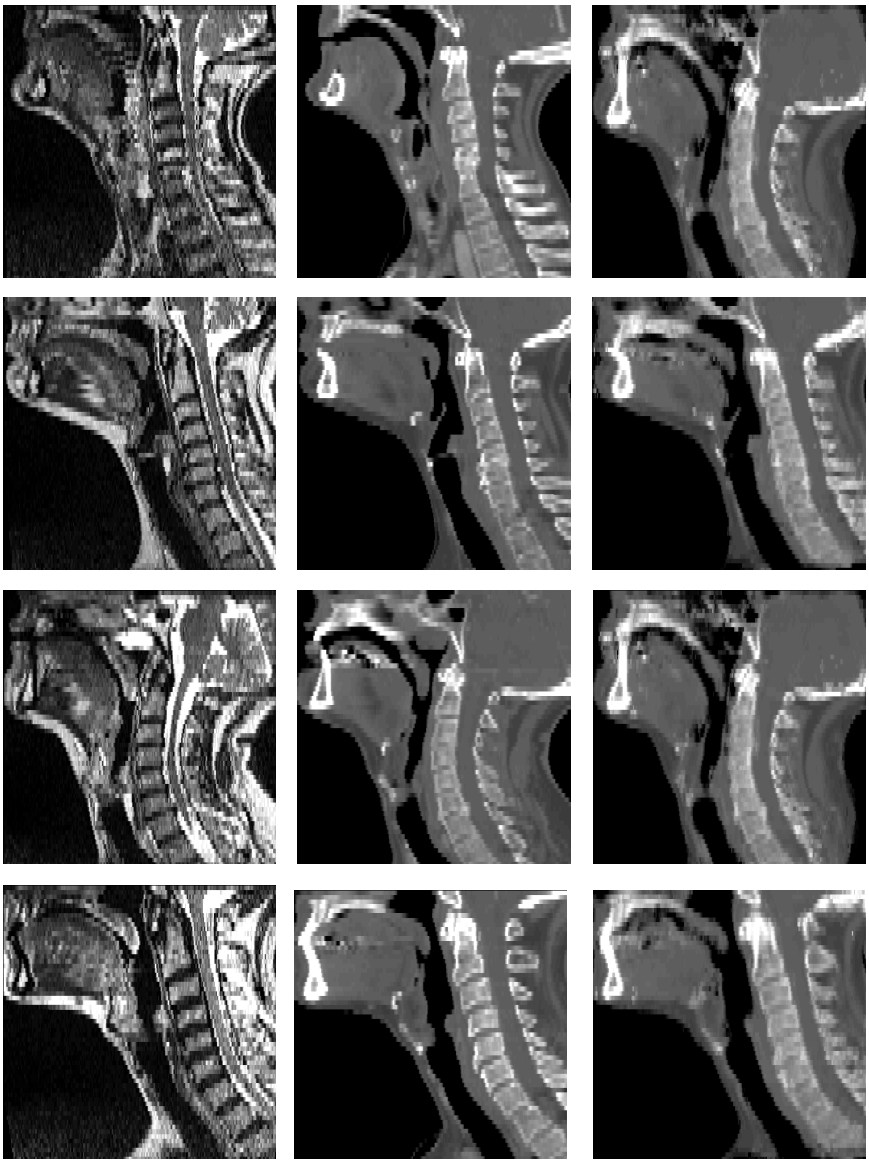

Figure 5.8: Dose calculation was done on 4 different patients. For each patient, dose calculation was estimated based on the individual original IMRT plan. Each row represents a patient. Left column: MR image. Middle column: real CT image. Left column: synthetic CT image.

Region of interest	Method	Pass percentage (%)
Treatment FOV	D_S	92.52 ($\pm$ 2.62)
	D_{Bulk}	90.38 ($\pm$ 6.80)
95% isodose volume	D_S	98.53 ($\pm$ 0.92)
	D_{Bulk}	82.94 ($\pm$ 12.88)

Table 5.1: Percentage of voxels within the region where the dose difference between D_S/D_{Bulk} and D_R is smaller than 2% of the prescribed dose.

based on two different anatomical regions: the treatment FOV (i.e the region of the body that receives 10% of the prescribed dose) and the 95% isodose volume (i.e the region of the body that receives 95% of the prescribed dose). Table 5.1 presents the percentage of voxels within the region where the dose difference is smaller than 2% of the prescribed dose. For both regions, D_S displayed higher similarities than D_{Bulk} when compared to D_R.

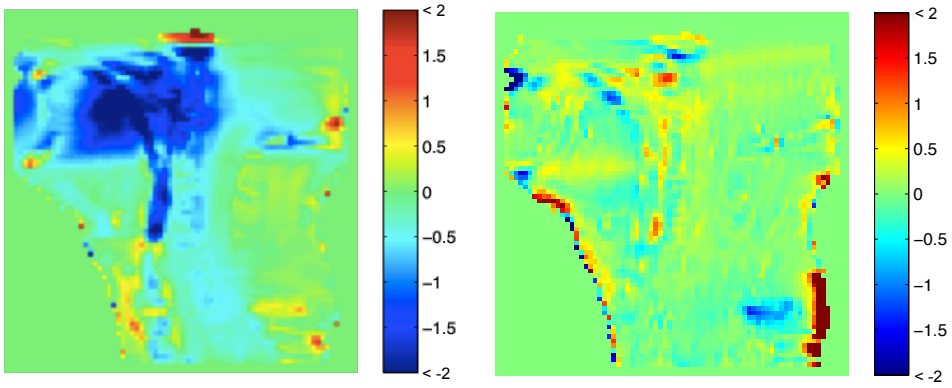

Figure 5.9: Absolute dose difference (Gy) between D_{Bulk}/D_S and D_R. Left bulk-assigned CT image. Right synthetic CT image. More dose is delivered to tissue when using $Bulk^{CT}$ compared to R^{CT}.

DVHs of the brainstem, spinal canal, left parotid, and right parotid are presented in Figure 5.10. DVHs from synthetic CT images agreed with those from real CT images. In constrast, DVHs from bulk assigned CT images deviated toward higher doses than the other DVHs. This was expected since the value of water was assigned to the bulk assigned CT images which has a lower CT number than tissues in real CT images, particularly at bony structures, which would result in less attenuated beams and, in turn, a higher dose distribution.

5.5 Discussion

MR imaging often provides improved contrast resolution between different types of tissues compared to CT imaging. Nevertheless, MR imaging is usually not used as a stand-alone in radiotherapy due to several technical limitations, one of them being the lack of electron density information for dose calculation. In this study, a novel method is presented for generating a synthetic CT image from an MR image to obtain an electron density map used for dose calculation. The method relies on an atlas-based propagation and integrates a morphological similarity measure and an atlas ranking scheme.

Results showed that S^{CT} images had good visual similarities with the R^{CT} images. The MAE estimated between the S^{CT} and the R^{CT} images is significantly smaller compared to the propagation of the single best atlas. However, discrepancies between the S^{CT} and the R^{CT} images could be observed around the bone/tissue and air/tissue interfaces where non-rigid registration in those area was not accurate. Results from dosimetry demonstrated that the S^{CT} images were better than the $Bulk^{CT}$ images showing higher pass percentage for different regions of interest. Dose distribution calculated on S^{CT} images were in close agreement with the one calculated on R^{CT}. In addition, DVHs of OARs from S^{CT} closely matched those from R^{CT}. Overall, this study shows promising results for the use of synthetic CT images in radiotherapy treatment planning.

A potential drawback of atlas-based methods for radiotherapy treatment planning is that image registration is associated with geometric uncertainties. In this study, registration problems were compensated by the local atlas selection and ranking steps. However, errors associated with atlas deformation can be significant if severe anatomical abnormalities are present in the target image. If the size of the

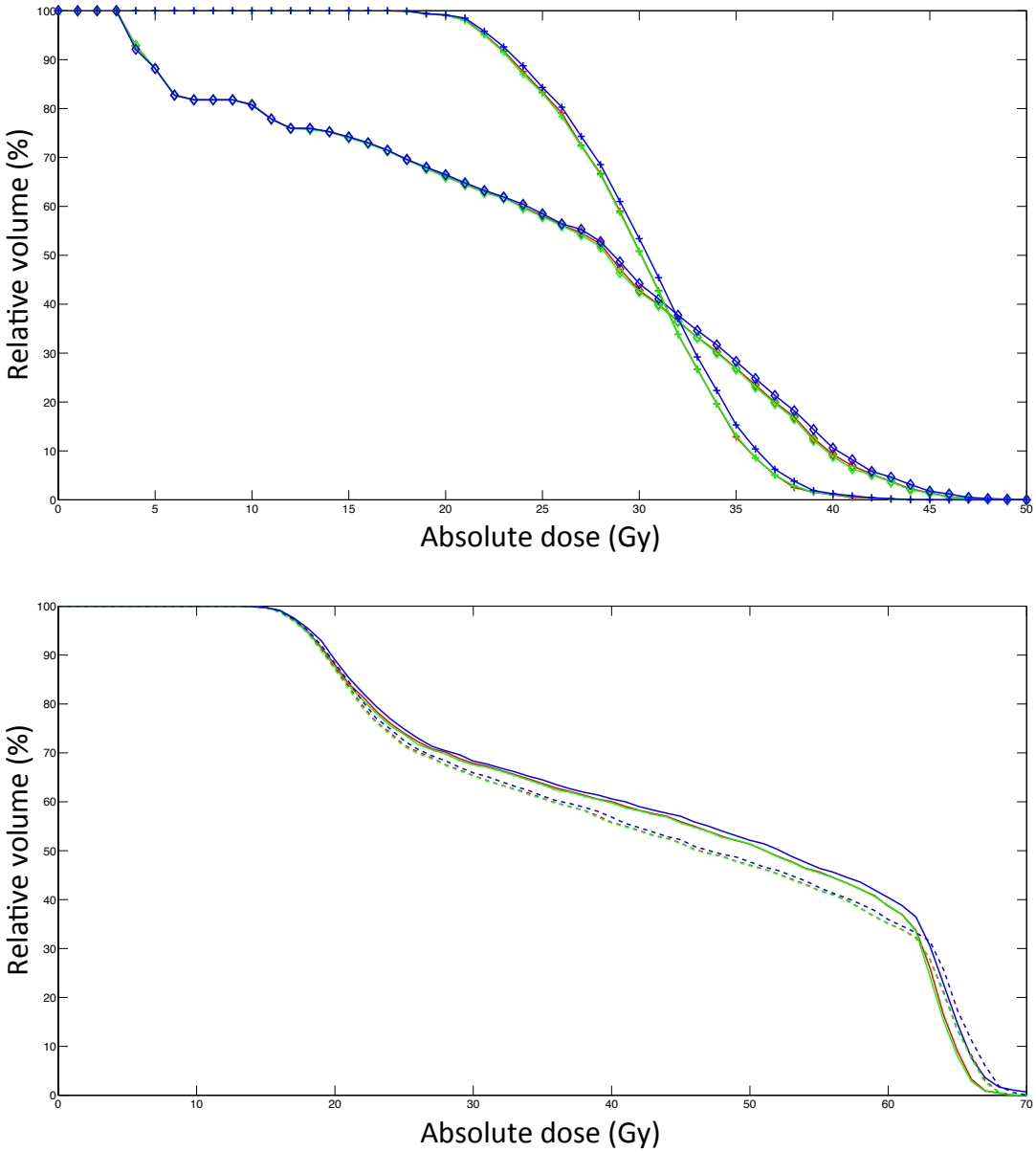

Figure 5.10: DVH for different OARs using dose distribution from real CT image (red), from synthetic CT image (green), and from bulk assigned CT image (blue). Top: brainstem (diamond lines) and spinal canal (cross lines). Bottom: left parotid (dashed lines) and right parotid (continuous lines).

dataset is large enough to cover a wide range of the population, the non-rigid registration should be able to capture those abnormalities. Further work will need to be completed on measuring the accuracy of the deformation maps and improving registration algorithms. One source of errors in the registration algorithm used in this study is the inherent deformation of bony elements, which physically can only move rigidly. A possible way to improve the proposed method could be to use poly-affine and poly-smooth registration algorithms (Arsigny et al., 2003).

Errors occurring in the generation of the synthetic CT images can be due to the intra-subject registration uncertainties between CT and MR images or the inter-subject registration uncertainties between MR and MR images. The fusion strategy can also contributes to those errors. Multiple hypothesis can be formulated: are dissimilar atlases being included when they should not be? Are similar atlases not being weighted high enough? Are there no atlases which are good enough? How does different fields of view and missing data from some atlases influence the fusion algorithm? Further research will need to be done in order to answer those questions.

The main limitation in this study is the small field of view on MR images. At the time of image acquisition, there were no clinical reasons for imaging the patients in treatment position with a larger field of view on the MR scan. However this limitation should not affect the conclusion from this study. The technique could, in theory, be applied to other body parts as long as the morphological variability is represented in the database and the registration between MR images is sufficiently accurate. Future work will include additional clinical validation by performing dose calculation on the full dataset, and making the method more robust by improving the registration algorithm as well as the fusion strategy.

5.6 Conclusion

In this chapter, the feasibility of using synthetic CT images for treatment planning of head and neck cancer generated from multiple deformed CT/MR atlases has been demonstrated. Synthetic CT images showed high similarities with the real CT images and the calculated doses agreed well with those based on real CT images. Larger scale studies need to be done in order to further validate the accuracy of synthetic CT images compared to real CT images for the broad workflow of radiotherapy applications in the head and neck. To conclude, the results obtained in this Chapter 5 along with the ones in Chapter 4 further support the use of atlas-based algorithms in radiotherapy treatment planning.

Chapter 6

General Conclusions

6.1 Summary

Radiotherapy treatment of head and neck cancer requires the delineation of OARs, a time-consuming and labour intensive process when performed manually. Accurate segmentation is critical in order to minimize radiation doses during treatment. This consequently improves life expectancy and reduces any negative impact on quality of life. In this thesis, I demonstrated that atlas-based method can produce clinically acceptable segmentations and reduce clinicians manual labour, helping them to focus on other aspects of patient's treatment. MR imaging is frequently used in radiotherapy planning as it has superior soft tissue contrast over that of CT imaging. However, it has not been clinically used alone as it does not provide electron density of tissues, rendering direct dose calculation on MR scans an impossible task to perform. CT imaging remains the main modality for radiotherapy planning because of a lack of correlation between MR intensities and electron density information. In this work, I showed that dose calculations based on synthetic CT images generated through an atlas-based method were in close agreement with full density CT-based plans. The proposed method in this thesis provides the necessary tools to MR imaging-based treatment planning feasible.

In Chapter 3, I proposed a new atlas-based segmentation method based on the out-of-sample property of manifold learning. The method is computationally fast and scalable making it suitable for segmenting large datasets of images acquired during radiotherapy planning. It was demonstrated that this method produces robust and accurate segmentation comparable to state-of-the-art methods. The results showed that selection of atlases with manifold learning is beneficial in the framework of multi-atlas based segmentation. The optimal accuracy can be found by fine tuning the manifold learning process.

In Chapter 4, I demonstrated that atlas-based method can produce segmentations graded as good as or better than manual contours with a rate of 83% in the context of radiotherapy planning and decrease manual labour. The reduction in segmentation time was on average 77%. In addition, I showed that the Dice similarity coefficient (DSC) does not reliably reflect the clinical acceptability of an automatic segmentation. Although a high DSC should guarantee clinical acceptability, a lower DSC does not necessarily mean that the segmentation produced by the proposed method was not clinically useful. The dataset used contained a variety of cases including some with bulky tumors, and results with the proposed method were still comparable to manual contouring across the cohort, demonstrating its robustness. In

any case, it was concluded that automatic segmentations should always be checked and corrected if necessary by an expert before planning.

In Chapter 5, I showed that synthetic CT images generated from MR scans with an atlas-based method can be used for radiotherapy planning. Fusion of multiple atlases using a local similarity measure and an atlas ranking scheme resulted in a synthetic CT image similar to the real CT image. The reported mean absolute error was 131.8 ($\pm$31.5) HU for this method, significantly lower than using the propagation of a single best atlas. The distribution of CT number in real and synthetic CT images were in close agreement as well as the dose volumes histograms generated from real and synthetic images. In addition, using a dose difference with a constraint of 2% prescribed dose, the pass percentage on the 95% isodose volumes was 98.58% ($\pm$0.92) when comparing dose distribution from real and synthetic CT image. Imaging cost during treatment can be reduced by using a single modality instead of multiple modalities, and I demonstrated the feasibility of MR imaging-based treatment planning. Manually delineated volumes on MR images have been shown to have lower inter-observer variability and are smaller that those on CT images. Extra margins added to account for delineation uncertainties could be reduced by using MR imaging, resulting in less tissues irradiated and a reduction in treatment toxicity. Being able to generate synthetic CT images could be very useful for adaptive and image guided radiotherapy using the new MR-LINACs that are currently being developed.

6.2 Future work

The future of imaging in radiotherapy planning is promising, and advances in technologies will contribute to a better definition of target volumes and organs at risk. This will in turn enable an increase in the precision of the calculation of radiation dose to the tumor, a reduction in toxicity and an improvement in treatment outcome. In addition, several solutions are currently implemented in order to allow MR imaging-based treatment planning to become a reality. This may ultimately lead to the elimination of CT imaging which has been the foundation of treatment planning for the past four decades. This will have several beneficial implications including a lower overall treatment cost and a reduced X-ray exposure.

The segmentation of target volumes has not been addressed in this thesis and remains the first future work to be done. Atlas-based delineation of tumours is very challenging due to large variability. As discussed in Chapter 4, atlas-based segmentation is highly dependent on the similarity between the underlying atlas and the patient. It performs well when the shape of the target is well represented in the dataset of atlases. However tumors have no predefined shape. This especially affect inter-patient registration accuracy. As demonstrated in this thesis, clinical evaluations of automatic segmentations still reveal the need for manual editing of automatic contours. The definition of a dataset with appropriate atlases remains an open question. In particular, for anatomical structures outside the brain, atlas datasets often fail to include the whole spectrum of variations. Presently no consensus exists on inclusion/exclusion rules or dataset size. In my opinion, atlas-based segmentation can benefit from customization of datasets. In this respect, an effort should be made to build application specific datasets and to develop rules for choosing the appropriate atlases. However, the final goal should be patient specific, on-the-fly atlas

selection that assures an appropriate feature matching between a target image and atlases.

As demonstrated in Chapter 4, clinical validation is required to ensure that automatic segmentation algorithms are suitable for radiotherapy planning. In this thesis, this was performed by a trained clinician on a 3-point grading scale. However, the development of an automatic clinical validation protocol will be beneficial. In the absence of segmentation ground truth, automatic segmentations are validated against manual contours. The Dice similarity coefficient (DSC) is commonly used for the validation of automatic segmentations in radiotherapy. When multiple manual contours are available for a given region of interest, the DSC is calculated for each manual contour individually. The multiple DSCs are then averaged. The drawback of such a validation measure is that it does not use the knowledge of expert agreements and disagreements which becomes important in the context of structure delineation for radiotherapy planning. The second research direction that I propose is the development of a method for assessing the quality of segmentations used in radiotherapy treatment planning that incorporates the level of agreement between several raters.

In Chapter 5, I demonstrated that treatment planning using MR images is feasible making the acquisition of CT scan unnecessary in the radiotherapy workflow. However, the proposed method was only tested on images with small field of view. A third future work is a follow up on my work with further clinical validation of the synthetic CT images, using MR images with field of view covering full anatomy of the head and neck region. This could include investigating if the geometrical accuracy is good enough to generate satisfactory digitally reconstructed radiograph for patient alignment and investigating if the dosimetric deviations are significant compared to a CT based plan.

Appendix A

Open Software Effort

Open source is a development approach that promotes transparency and promises more quality, reliability and flexibility in the production and testing of software. Due to his open nature, most licenses allow anyone to contribute, understand, re-factor and reuse the code with no restrictions. As a supporter of this approach, all the code developed during my PhD is available under a Berkeley Software Distribution (BSD) licence. With a BSD licence, redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code or binaries must retain the copyright notices, the list of conditions and a disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of the organisation nor the names of its distributors may be used to endorse or promote products derived from this software without specific prior written permission.

A.1 Manifold learning software package

The manifold learning algorithms detailed in Chapter 2 have been implemented in C++: Principal Component Analysis (PCA), Isomap (ISO), Locally linear embedding (LLE) and Laplacian Eigenmaps (LEM). A list of images are taken as an input, as well as a mask defining a region of interest. A text file containing the coordinates of the nifty images in the low-dimensional space is recorded as an output. Examples of command line for each algorithm are as follows:

- `run_pca -in image_*.nii -mask mask.nii -out pca_mapping.txt`
- `run_iso -in image_*.nii -mask mask.nii -k 10 -out iso_mapping.txt`
- `run_lle -in image_*.nii -mask mask.nii -k 10 -reg 0.001 -out lle_mapping.txt`
- `run_lem -in image_*.nii -mask mask.nii -k 10 -rho 0.1 -out lem_mapping.txt`

where the description of the parameters are as follows:

- `-in`: the **input** parameter to specify a list of images, i.e: `image_001.nii image_002.nii image_003.nii ...`

- `-mask`: the **input** parameter to specify a binary mask in the space of the images above. The mask specifies the region of interest in the high-dimensional space.
- `-k`: the **input** parameter to specify the number of neighbours to built the connected graph in the high-dimensional space. Must be an integer.
- `-reg`: the **input** parameter to specify a regularization term. When the local covariance C is not full rank, it should be regularized with a small constant of order $Trace(C) * 10^{-3}$. Only necessary when using LLE.
- `-rho`: the **input** parameter to specify the standard deviation of the Gaussian kernel. Only necessary when using LEM.
- `-out`: the **ouput** text file containing the coordinates of the images in the low-dimensional space.

There are 4 main libraries _pca, _iso, _lle, and _lem making up the main application programming interface (API) for external linkage. Each one of these libraries defines a C++ object that defines, sets and runs the model and the necessary variables.

Bibliography

Acosta, O., Simon, A., Monge, F., Commandeur, F., Bassirou, C., Cazoulat, G., De Crevoisier, R., Haigron, P., 2011. Evaluation of multi-atlas-based segmentation of ct scans in prostate cancer radiotherapy. In: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. IEEE, pp. 1966–1969.

Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage 46 (3), 726–738.

Aljabar, P., Rueckert, D., Crum, W. R., 2008. Automated morphological analysis of magnetic resonance brain imaging using spectral analysis. Neuroimage 43 (2), 225–235.

Aljabar, P., Wolz, R., Srinivasan, L., Counsell, S., Boardman, J. P., Murgasova, M., Doria, V., Rutherford, M. A., Edwards, A. D., Hajnal, J. V., et al., 2010. Combining morphological information in a manifold learning framework: application to neonatal mri. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010. Springer, pp. 1–8.

Arsigny, V., Pennec, X., Ayache, N., 2003. Polyrigid and polyaffine transformations: A new class of diffeomorphisms for locally rigid or affine registration. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003. Springer, pp. 829–837.

Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C., 2009. Combination strategies in multi-atlas image segmentation: Application to brain mr data. Medical Imaging, IEEE Transactions on 28 (8), 1266–1277.

Asman, A. J., Landman, B. A., 2012. Non-local staple: An intensity-driven multi-atlas rater model. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer, pp. 426–434.

Avants, B., Gee, J. C., 2004. Geodesic estimation for large deformation anatomical shape averaging and interpolation. Neuroimage 23, S139–S150.

Awate, S. P., Zhu, P., Whitaker, R. T., 2012. How many templates does it take for a good segmentation?: error analysis in multiatlas segmentation as a function of database size. In: Multimodal Brain Image Analysis. Springer, pp. 103–114.

Barnes, J., Boyes, R., Lewis, E., Schott, J., Frost, C., Scahill, R., Fox, N., 2007. Automatic calculation of hippocampal atrophy rates using a hippocampal template and the boundary shift integral. Neurobiology of aging 28 (11), 1657–1663.

Barnes, J., Foster, J., Boyes, R., Pepple, T., Moore, E., Schott, J., Frost, C., Scahill, R., Fox, N., 2008a. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. Neuroimage 40 (4), 1655–1671.

Barnes, J., Scahill, R., Frost, C., Schott, J., Rossor, M., Fox, N., 2008b. Increased hippocampal atrophy rates in ad over 6 months using serial mr imaging. Neurobiology of aging 29 (8), 1199–1203.

Beg, M. F., Miller, M. I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision 61 (2), 139–157.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation 15 (6), 1373–1396.

Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M., 2004. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. Advances in neural information processing systems 16, 177–184.

Blezek, D. J., Miller, J. V., 2007. Atlas stratification. Medical Image Analysis 11 (5), 443–457.

Brandt, R., Rohlfing, T., Rybak, J., Krofczik, S., Maye, A., Westerhoff, M., Hege, H.-C., Menzel, R., 2005. Three-dimensional average-shape atlas of the honeybee brain and its applications. Journal of Comparative Neurology 492 (1), 1–19.

Burgos, N., Cardoso, J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C., Schott, J., et al., 2013. Attenuation correction synthesis for hybrid pet/mri scanners: application to brain studies. IEEE Trans Med Imaging (under review).

Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N., 2003. Iconic feature based nonrigid registration: the pasha algorithm. Computer vision and image understanding 89 (2), 272–298.

Cardoso, M. J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N. C., Ourselin, S., 2013. Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Medical image analysis.

Cardoso, M. J., Wolz, R., Modat, M., Fox, N. C., Rueckert, D., Ourselin, S., 2012. Geodesic information flows. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer, pp. 262–270.

Carmichael, O. T., Aizenstein, H. A., Davis, S. W., Becker, J. T., Thompson, P. M., Meltzer, C. C., Liu, Y., 2005. Atlas-based hippocampus segmentation in alzheimer's disease and mild cognitive impairment. Neuroimage 27 (4), 979–990.

Cefaro, G. A., Genovesi, D., Perez, C. A., 2013. Delineating organs at risk in radiation therapy.

Chao, K., Bhide, S., Chen, H., Asper, J., Bush, S., Franklin, G., Kavadi, V., Liengswangwong, V., Gordon, W., Raben, A., et al., 2007. Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach. International Journal of Radiation Oncology* Biology* Physics 68 (5), 1512–1521.

Christensen, G. E., Rabbitt, R. D., Miller, M. I., 1996. Deformable templates using large deformation kinematics. Image Processing, IEEE Transactions on 5 (10), 1435–1447.

Chupin, M., Chetelat, G., Lemieux, L., Dubois, B., Garnero, L., Benali, H., Eustache, F., Lehericy, S., Desgranges, B., Colliot, O., 2008. Fully automatic hippocampus segmentation discriminates between early alzheimers disease and normal aging. In: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on. IEEE, pp. 97–100.

Chupin, M., Mukuna-Bantumbakulu, A. R., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnéhun, S., Lemieux, L., Dubois, B., Garnero, L., 2007. Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with alzheimers disease. Neuroimage 34 (3), 996–1019.

Collins, D. L., Holmes, C. J., Peters, T. M., Evans, A. C., 1995. Automatic 3-d model-based neuroanatomical segmentation. Human brain mapping 3 (3), 190–208.

Collins, D. L., Neelin, P., Peters, T. M., Evans, A. C., 1994. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. Journal of computer assisted tomography 18 (2), 192–205.

Collins, D. L., Pruessner, J. C., 2010. Towards accurate, automatic segmentation of the hippocampus and amygdala from mri by augmenting animal with a template library and label fusion. NeuroImage 52 (4), 1355–1366.

Commowick, O., Grégoire, V., Malandain, G., 2008. Atlas-based delineation of lymph node levels in head and neck computed tomography images. Radiotherapy and Oncology 87 (2), 281–289.

Commowick, O., Malandain, G., 2007. Efficient selection of the most similar image in a database for critical structures segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007. Springer, pp. 203–210.

Commowick, O., Malandain, G., et al., 2006. Evaluation of atlas construction strategies in the context of radiotherapy planning. In: Proceedings of the SA2PM Workshop (From Statistical Atlases to Personalized Models).

Commowick, O., Warfield, S. K., Malandain, G., 2009. Using frankensteins creature paradigm to build a patient specific atlas. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009. Springer, pp. 993–1000.

Cox, T. F., Cox, M. A., 2010. Multidimensional scaling. CRC Press.

Daisne, J.-F., Blumhofer, A., 2013. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. Radiation Oncology 8 (1), 154.

Dearnaley, D. P., Khoo, V. S., Norman, A. R., Meyer, L., Nahum, A., Tait, D., Yarnold, J., Horwich, A., 1999. Comparison of radiation side-effects of conformal and conventional radiotherapy in prostate cancer: a randomised trial. The Lancet 353 (9149), 267–272.

Dice, L. R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Dijkstra, E. W., 1959. A note on two problems in connexion with graphs. Numerische mathematik 1 (1), 269–271.

Dong, L., Boyer, A. L., 1995. An image correlation procedure for digitally reconstructed radiographs and electronic portal images. International Journal of Radiation Oncology* Biology* Physics 33 (5), 1053–1060.

Doran, S. J., Charles-Edwards, L., Reinsberg, S. A., Leach, M. O., 2005. A complete distortion correction for mr images: I. gradient warp correction. Physics in medicine and biology 50 (7), 1343.

Dowling, J. A., Lambert, J., Parker, J., Salvado, O., Fripp, J., Capp, A., Wratten, C., Denham, J. W., Greer, P. B., 2012. An atlas-based electron density mapping method for magnetic resonance imaging (mri)-alone treatment planning and adaptive mri-based prostate radiation therapy. International Journal of Radiation Oncology* Biology* Physics 83 (1), e5–e11.

Duchesne, S., Pruessner, J., Collins, D., 2002. Appearance-based segmentation of medial temporal lobe structures. Neuroimage 17 (2), 515–531.

Edge, S. B., Compton, C. C., 2010. The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. Annals of surgical oncology 17 (6), 1471–1474.

Evans, P. M., 2008. Anatomical imaging for radiotherapy. Physics in medicine and biology 53 (12), R151.

Fiorino, C., Reni, M., Bolognesi, A., Cattaneo, G. M., Calandrino, R., 1998. Intra-and inter-observer variability in contouring prostate and seminal vesicles: implications for conformal treatment planning. Radiotherapy and Oncology 47 (3), 285–292.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355.

Fletcher, P. T., Venkatasubramanian, S., Joshi, S., 2009. The geometric median on riemannian manifolds with application to robust atlas estimation. NeuroImage 45 (1), S143–S152.

Floyd, R. W., 1962. Algorithm 97: shortest path. Communications of the ACM 5 (6), 345.

Fraass, B., Doppke, K., Hunt, M., Kutcher, G., Starkschall, G., Stern, R., Van Dyke, J., 1998. American association of physicists in medicine radiation therapy committee task group 53: quality assurance for clinical radiotherapy treatment planning. Medical physics 25 (10), 1773–1829.

Freeborough, P. A., Fox, N. C., 1998. Modeling brain deformations in alzheimer disease by fluid registration of serial 3d mr images. Journal of computer assisted tomography 22 (5), 838–843.

Gee, J. C., Reivich, M., Bajcsy, R., 1993. Elastically deforming 3d atlas to match anatomical brain images. Journal of computer assisted tomography 17 (2), 225–236.

Geets, X., Daisne, J.-F., Arcangeli, S., Coche, E., Poel, M. D., Duprez, T., Nardella, G., Grégoire, V., 2005. Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between ct-scan and mri. Radiotherapy and oncology 77 (1), 25–31.

Georg, M., Souvenir, R., Hope, A., Pless, R., 2008. Manifold learning for 4d ct reconstruction of the lung. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on. IEEE, pp. 1–8.

Gerber, S., Tasdizen, T., Thomas Fletcher, P., Joshi, S., Whitaker, R., 2010. Manifold modeling for brain population analysis. Medical image analysis 14 (5), 643–653.

Gousias, I. S., Rueckert, D., Heckemann, R. A., Dyet, L. E., Boardman, J. P., Edwards, A. D., Hammers, A., 2008. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. Neuroimage 40 (2), 672–684.

Grégoire, V., Eisbruch, A., Hamoir, M., Levendag, P., 2006. Proposal for the delineation of the nodal ctv in the node-positive and the post-operative neck. Radiotherapy and oncology 79 (1), 15–20.

Grégoire, V., Levendag, P., Ang, K. K., Bernier, J., Braaksma, M., Budach, V., Chao, C., Coche, E., Cooper, J. S., Cosnard, G., et al., 2003. Ct-based delineation of lymph node levels and related ctvs in the node-negative neck: Dahanca, eortc, gortec, ncic, rtog consensus guidelines. Radiotherapy and oncology 69 (3), 227–236.

Grosu, A.-L., 2006. Stereotactic radiotherapy/radiosurgery. In: New Technologies in Radiation Oncology. Springer, pp. 267–276.

Guimond, A., Meunier, J., Thirion, J.-P., 2000. Average brain models: A convergence study. Computer vision and image understanding 77 (2), 192–210.

Gunter, J., Bernstein, M., Borowski, B., Felmlee, J., Blezek, D., Mallozzi, R., Levy, J., Schuff, N., Jack, C., 2006. Validation testing of the mri calibration phantom for the alzheimer's disease neuroimaging initiative study. International Society for Magnetic Resonance in Medicine, Seattle, WA.

Hamm, J., Ye, D. H., Verma, R., Davatzikos, C., 2010. Gram: A framework for geodesic registration on anatomical manifolds. Medical image analysis 14 (5), 633–642.

Han, X., Hoogeman, M. S., Levendag, P. C., Hibbard, L. S., Teguh, D. N., Voet, P., Cowen, A. C., Wolf, T. K., 2008. Atlas-based auto-segmentation of head and neck ct images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008. Springer, pp. 434–441.

Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., Hammers, A., 2006a. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. NeuroImage 33 (1), 115–126.

Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., Hammers, A., 2006b. Multiclassifier fusion in human brain mr segmentation: modelling convergence. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006. Springer, pp. 815–822.

Hoang Duc, A. K., Modat, M., Leung, K. K., Cardoso, M. J., Barnes, J., Kadir, T., Ourselin, S., Initiative, A. D. N., 2013. Using manifold learning for atlas selection in multi-atlas segmentation. PloS one 8 (8), e70059.

Hong, T., Tome, W., Chappell, R., Harari, P., 2004. Variations in target delineation for head and neck imrt: An international multi-institutional study. International Journal of Radiation Oncology* Biology* Physics 60 (1), S157–S158.

Hoppe, R., Phillips, T. L., Roach III, M., 2010. Leibel and Phillips Textbook of Radiation Oncology: Expert Consult. Elsevier Health Sciences.

Hu, S., Collins, D. L., 2007. Joint level-set shape modeling and appearance modeling for brain structure segmentation. NeuroImage 36 (3), 672–683.

Irani, S. R., Stagg, C. J., Schott, J. M., Rosenthal, C. R., Schneider, S. A., Pettingill, P., Pettingill, R., Waters, P., Thomas, A., Voets, N. L., et al., 2013. Faciobrachial dystonic seizures: the influence of immunotherapy on seizure control and prevention of cognitive impairment in a broadening phenotype. Brain 136 (10), 3151–3162.

Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M. A., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusionapplication to cardiac and aortic segmentation in ct scans. Medical Imaging, IEEE Transactions on 28 (7), 1000–1010.

Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. New phytologist 11 (2), 37–50.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L Whitwell, J., Ward, C., et al., 2008. The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging 27 (4), 685–691.

Jeanneret-Sozzi, W., Moeckli, R., Valley, J.-F., Zouhair, A., Ozsahin, E. M., Mirimanoff, R.-O., 2006. The reasons for discrepancies in target volume delineation. Strahlentherapie und Onkologie 182 (8), 450–457.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical image analysis 5 (2), 143–156.

Jolliffe, I., 2005. Principal component analysis. Wiley Online Library.

Jongen, C., Pluim, J. P., Nederkoorn, P. J., Viergever, M. A., Niessen, W. J., 2004. Construction and evaluation of an average ct brain image for inter-subject registration. Computers in biology and medicine 34 (8), 647–662.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., et al., 2006. Reliability in multi-site structural mri studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 30 (2), 436–443.

Karotki, A., Mah, K., Meijer, G., Meltsner, M., 2011. Comparison of bulk electron density and voxel-based electron density treatment planning. Journal of Applied Clinical Medical Physics 12 (4).

Klein, A., Hirsch, J., 2005. Mindboggle: a scatterbrained approach to automate brain labeling. NeuroImage 24 (2), 261–280.

Klein, S., van der Heide, U. A., Lips, I. M., van Vulpen, M., Staring, M., Pluim, J. P., 2008. Automatic segmentation of the prostate in 3d mr images by atlas matching using localized mutual information. Medical physics 35 (4), 1407–1417.

Kruser, T. J., Bradley, K. A., Bentzen, S. M., Anderson, B. M., Gondi, V., Khuntia, D., Perlman, S. B., Tome, W. A., Chappell, R. J., Walker, W. L., et al., 2009. The impact of hybrid pet-ct scan on overall oncologic management, with a focus on radiotherapy planning: a prospective, blinded study. Technology in cancer research & treatment 8 (2), 149–158.

Langerak, T. R., Berendsen, F. F., Van der Heide, U. A., Kotte, A. N. T. J., Pluim, J. P. W., 2013. Multi-atlas based segmentation with preregistration atlas selection. Medical Physics 40 (9).

Lee, Y. K., Bollet, M., Charles-Edwards, G., Flower, M. A., Leach, M. O., McNair, H., Moore, E., Rowbottom, C., Webb, S., 2003. Radiotherapy treatment planning of prostate cancer using magnetic resonance imaging alone. Radiotherapy and oncology 66 (2), 203–216.

Leksell, L., 1983. Stereotactic radiosurgery. Journal of Neurology, Neurosurgery & Psychiatry 46 (9), 797–803.

Leung, K. K., Barnes, J., Ridgway, G. R., Bartlett, J. W., Clarkson, M. J., Macdonald, K., Schuff, N., Fox, N. C., Ourselin, S., 2010. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and alzheimer's disease. Neuroimage 51 (4), 1345–1359.

Lewis, J., 1995. Fast normalized cross-correlation. In: Vision interface. Vol. 10. pp. 120–123.

Li, B., Christensen, G. E., Hoffman, E. A., McLennan, G., Reinhardt, J. M., 2003. Establishing a normative atlas of the human lung: intersubject warping and registration of volumetric ct images. Academic Radiology 10 (3), 255–265.

Lötjönen, J. M., Wolz, R., Koikkalainen, J. R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 49 (3), 2352–2365.

Ma, D., Cardoso, M. J., Modat, M., Powell, N., Wells, J., Holmes, H., Wiseman, F., Tybulewicz, V., Fisher, E., Lythgoe, M. F., et al., 2014. Automatic structural parcellation of mouse brain mri using multi-atlas label fusion. PloS one 9 (1), e86576.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. Medical Imaging, IEEE Transactions on 16 (2), 187–198.

Marshall, H. R., Patrick, J., Laidley, D., Prato, F. S., Butler, J., Théberge, J., Thompson, R. T., Stodilka, R. Z., 2013. Description and assessment of a registration-based approach to include bones for attenuation correction of whole-body pet/mri. Medical physics 40 (8), 082509.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E. M., 1984. Clinical diagnosis of alzheimer's disease report of the nincds-adrda work group* under the auspices of department of health and human services task force on alzheimer's disease. Neurology 34 (7), 939–939.

Meredith, W., 1984. 40 years of development in radiotherapy. Physics in medicine and biology 29 (2), 115.

Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine 98 (3), 278–284.

Morra, J. H., Tu, Z., Apostolova, L. G., Green, A. E., Avedissian, C., Madsen, S. K., Parikshak, N., Hua, X., Toga, A. W., Jack Jr, C. R., et al., 2008. Validation of a fully automated 3d hippocampal segmentation method using subjects with alzheimer's disease mild cognitive impairment, and elderly controls. Neuroimage 43 (1), 59–68.

Narayana, P., Brey, W., Kulkarni, M., Sievenpiper, C., 1988. Compensation for surface coil sensitivity variation in magnetic resonance imaging. Magnetic resonance imaging 6 (3), 271–274.

Nelms, B. E., Tomé, W. A., Robinson, G., Wheeler, J., 2012. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. International Journal of Radiation Oncology* Biology* Physics 82 (1), 368–378.

Newbold, K., Partridge, M., Cook, G., Sohaib, S., Charles-Edwards, E., Rhys-Evans, P., Harrington, K., Nutting, C., 2006. Advanced imaging applied to radiotherapy planning in head and neck cancer: a clinical review. Advanced imaging 79 (943).

Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N., 2001. Reconstructing a 3d structure from serial histological sections. Image and vision computing 19 (1), 25–31.

Parker, R., Hobday, P. A., Cassell, K., 1979. The direct use of ct numbers in radiotherapy dosage calculations for inhomogeneous media. Physics in medicine and biology 24 (4), 802.

Pless, R., 2004. Differential structure in non-linear image embedding functions. In: Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on. IEEE, pp. 10–10.

Pohl, K. M., Bouix, S., Shenton, M. E., Grimson, W. E. L., Kikinis, R., 2007. Automatic segmentation using non-rigid registration. In: Medical image computing and computer-assisted intervention: MICCAI... International Conference on Medical Image Computing and Computer-Assisted Intervention. Vol. 26. NIH Public Access, p. 1201.

Powell, S., Magnotta, V. A., Johnson, H., Jammalamadaka, V. K., Pierson, R., Andreasen, N. C., 2008. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. Neuroimage 39 (1), 238–247.

Ramus, L., Commowick, O., Malandain, G., 2010. Construction of patient specific atlases from locally most similar anatomical pieces. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010. Springer, pp. 155–162.

Rasch, C., Barillot, I., Remeijer, P., Touw, A., van Herk, M., Lebesque, J. V., 1999. Definition of the prostate in ct and mri: a multi-observer study. International Journal of Radiation Oncology* Biology* Physics 43 (1), 57–66.

Rasch, C., Keus, R., Pameijer, F. A., Koops, W., de Ru, V., Muller, S., Touw, A., Bartelink, H., van Herk, M., Lebesque, J. V., 1997. The potential impact of ct-mri matching on tumor volume delineation in advanced head and neck cancer. International Journal of Radiation Oncology* Biology* Physics 39 (4), 841–848.

Roach III, M., Faillace-Akazawa, P., Malfatti, C., Holland, J., Hricak, H., 1996. Prostate volumes defined by magnetic resonance imaging and computerized tomographic scans for three-dimensional conformal radiotherapy. International Journal of Radiation Oncology* Biology* Physics 35 (5), 1011–1018.

Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr, C. R., 2004a. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21 (4), 1428–1442.

Rohlfing, T., Brandt, R., Menzel, R., Russakoff, D. B., Maurer Jr, C. R., 2005. Quo vadis, atlas-based segmentation? In: Handbook of Biomedical Image Analysis. Springer, pp. 435–486.

Rohlfing, T., Russakoff, D. B., Maurer Jr, C. R., 2004b. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. Medical Imaging, IEEE Transactions on 23 (8), 983–994.

Roweis, S. T., Saul, L. K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Rueckert, D., Schnabel, J. A., 2011. Medical image registration. In: Biomedical Image Processing. Springer, pp. 131–154.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., Hawkes, D. J., 1999. Nonrigid registration using free-form deformations: application to breast mr images. Medical Imaging, IEEE Transactions on 18 (8), 712–721.

Sabuncu, M. R., Balci, S. K., Golland, P., 2008. Discovering modes of an image population through mixture modeling. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008. Springer, pp. 381–389.

Sabuncu, M. R., Yeo, B. T., Van Leemput, K., Fischl, B., Golland, P., 2010. A generative model for image segmentation based on label fusion. Medical Imaging, IEEE Transactions on 29 (10), 1714–1729.

Scheltens, P., Pasquier, F., Weerts, J., Barkhof, F., Leys, D., 1997. Qualitative assessment of cerebral atrophy on mri: inter-and intra-observer reproducibility in dementia and normal aging. European neurology 37 (2), 95–99.

Sdika, M., 2010. Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. Medical Image Analysis 14 (2), 219–226.

Seco, J., Evans, P., 2006. Assessing the effect of electron density in photon dose calculations. Medical physics 33 (2), 540–552.

Shen, D., Davatzikos, C., 2002. Hammer: hierarchical attribute matching mechanism for elastic registration. Medical Imaging, IEEE Transactions on 21 (11), 1421–1439.

Shen, D., Moffat, S., Resnick, S. M., Davatzikos, C., 2002. Measuring size and shape of the hippocampus in mr images using a deformable shape model. Neuroimage 15 (2), 422–434.

Sjöberg, C., Lundmark, M., Granberg, C., Johansson, S., Ahnesjö, A., Montelius, A., 2013. Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients. Radiation Oncology 8 (1), 229.

Skrzyński, W., Zielińska-Dabrowska, S., Wachowicz, M., Ślusarczyk-Kacprzyk, W., Kukołowicz, P. F., Bulski, W., 2010. Computed tomography as a source of electron density information for radiation treatment planning. Strahlentherapie und Onkologie 186 (6), 327–333.

Sled, J. G., Zijdenbos, A. P., Evans, A. C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. Medical Imaging, IEEE Transactions on 17 (1), 87–97.

Sluimer, I., Prokop, M., Van Ginneken, B., 2005. Toward automated segmentation of the pathological lung in ct. Medical Imaging, IEEE Transactions on 24 (8), 1025–1038.

Snow, G., Annyas, A., Slooten, E. v., Bartelink, H., Hart, A., 1982. Prognostic factors of neck node metastasis. Clinical Otolaryngology & Allied Sciences 7 (3), 185–192.

Som, P. M., Curtin, H. D., Mancuso, A. A., 2000. Imaging-based nodal classification for evaluation of neck metastatic adenopathy. American Journal of Roentgenology 174 (3), 837–844.

Souvenir, R., Pless, R., 2005. Isomap and nonparametric models of image deformation. In: Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on. Vol. 2. IEEE, pp. 195–200.

Souvenir, R., Pless, R., 2007. Image distance functions for manifold learning. Image and Vision Computing 25 (3), 365–373.

Stanescu, T., Jans, H. S., Wachowicz, K., Fallone, B. G., 2010. Investigation of a 3d system distortion correction method for mr images. Journal of Applied Clinical Medical Physics 11 (1).

Stapleford, L. J., Lawson, J. D., Perkins, C., Edelman, S., Davis, L., McDonald, M. W., Waller, A., Schreibmann, E., Fox, T., 2010. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. International Journal of Radiation Oncology* Biology* Physics 77 (3), 959–966.

Teguh, D. N., Levendag, P. C., Voet, P. W., Al-Mamgani, A., Han, X., Wolf, T. K., Hibbard, L. S., Nowak, P., Akhiat, H., Dirkx, M. L., et al., 2011. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. International Journal of Radiation Oncology* Biology* Physics 81 (4), 950–957.

Tenenbaum, J. B., De Silva, V., Langford, J. C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Thirion, J.-P., 1998. Image matching as a diffusion process: an analogy with maxwell's demons. Medical image analysis 2 (3), 243–260.

Uh, J., Merchant, T., Hua, C., 2013. Th-c-wab-11: Mri-based treatment planning with pseudo ct generated through atlas registration. Medical Physics 40 (6), 538–538.

Uh, J., Merchant, T. E., Li, Y., Li, X., Hua, C., 2014. Mri-based treatment planning with pseudo ct generated through atlas registration. Medical physics 41 (5), 051711.

van der Lijn, F., den Heijer, T., Breteler, M., Niessen, W. J., 2008. Hippocampus segmentation in mr images using atlas registration, voxel classification, and graph cuts. Neuroimage 43 (4), 708–720.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage 45 (1), S61–S72.

Vokes, E. E., Weichselbaum, R. R., Lippman, S. M., Hong, W. K., 1993. Head and neck cancer. New England Journal of Medicine 328 (3), 184–194.

Wachinger, C., Navab, N., 2010. Manifold learning for multi-modal image registration. In: BMVC. pp. 1–12.

Wang, H., Suh, J., Das, S., Pluta, J., Craige, C., Yushkevich, P., 2013. Multi-atlas segmentation with joint label fusion. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (3), 611–623.

Warfield, S. K., Zou, K. H., Wells, W. M., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. Medical Imaging, IEEE Transactions on 23 (7), 903–921.

Webb, S., 2001. Intensity-modulated radiation therapy. CRC Press.

Weiss, E., Hess, C. F., 2003. The impact of gross tumor volume (gtv) and clinical target volume (ctv) definition on the total accuracy in radiotherapy. Strahlentherapie und Onkologie 179 (1), 21–30.

Wells III, W. M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Medical image analysis 1 (1), 35–51.

Wolz, R., Aljabar, P., Hajnal, J. V., Hammers, A., Rueckert, D., 2010a. Leap: learning embeddings for atlas propagation. NeuroImage 49 (2), 1316–1325.

Wolz, R., Aljabar, P., Hajnal, J. V., Rueckert, D., 2010b. Manifold learning for biomarker discovery in mr imaging. In: Machine Learning in Medical Imaging. Springer, pp. 116–123.

Wu, A., Lindner, G., Maitz, A., Kalend, A., Lunsford, L., Flickinger, J., Bloomer, W., 1990. Physics of gamma knife approach on convergent beams in stereotactic radiosurgery. International Journal of Radiation Oncology* Biology* Physics 18 (4), 941–949.

Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C. S., Aizenstein, H. J., 2007. Optimum template selection for atlas-based segmentation. NeuroImage 34 (4), 1612–1618.

Xing, L., Thorndyke, B., Schreibmann, E., Yang, Y., Li, T.-F., Kim, G.-Y., Luxton, G., Koong, A., 2006. Overview of image-guided radiation therapy. Medical Dosimetry 31 (2), 91–112.

Xu, L., Krzyzak, A., Suen, C. Y., 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. Systems, Man and Cybernetics, IEEE Transactions on 22 (3), 418–435.

Yang, J., Zhang, Y., Zhang, L., Dong, L., 2010. Automatic segmentation of parotids from ct scans using multiple atlases. Medical Image Analysis for the Clinic: A Grand Challenge, 323–330.

Young, A. V., Wortham, A., Wernick, I., Evans, A., Ennis, R. D., 2011. Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes. International Journal of Radiation Oncology* Biology* Physics 79 (3), 943–947.

Yushkevich, P. A., Wang, H., Pluta, J., Das, S. R., Craige, C., Avants, B. B., Weiner, M. W., Mueller, S., 2010. Nearly automatic segmentation of hippocampal subfields in in vivo focal t2-weighted mri. Neuroimage 53 (4), 1208–1224.

Zhang, L., Hoffman, E. A., Reinhardt, J. M., 2006a. Atlas-driven lung lobe segmentation in volumetric x-ray ct images. Medical Imaging, IEEE Transactions on 25 (1), 1–16.

Zhang, Q., Souvenir, R., Pless, R., 2006b. On manifold structure of cardiac mri data: Application to segmentation. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 1. IEEE, pp. 1092–1098.